Table of Contents
1 Introduction ........................................................................................................................................... 3
2 Knowledge Engineering Processes ........................................................................................................... 3
  2.1 Content Creation ......................................................................................................................................... 4
    2.1.1 Select Relevant Sentences ................................................................................................................... 5
    2.1.2 Paraphrase Sentences as Universal Truths ............................................................................................ 6
    2.1.3 Create and Review the Taxonomy .......................................................................................................... 12
    2.1.4 Concept Representation Planning ......................................................................................................... 16
    2.1.5 Concept Map Creation .......................................................................................................................... 17
  2.2 Methodology for Testing Knowledge ....................................................................................................... 18
    2.2.1 Testing During Knowledge Encoding .................................................................................................... 18
    2.2.2 Testing After Knowledge Encoding ....................................................................................................... 21
  2.3 Quality Assurance ................................................................................................................................. 23
    2.3.1 Concept Map Review by Biologists ........................................................................................................ 22
    2.3.2 Answer Output Review by Biologists .................................................................................................... 23
    2.3.3 Knowledge Representation Quality Analysis .......................................................................................... 23
  2.4 Reporting Progress ............................................................................................................................... 26
    2.4.1 Coverage Analysis .................................................................................................................................. 26
    2.4.2 Regression Testing .................................................................................................................................. 32
    2.4.3 Reporting Crashes .................................................................................................................................. 32
  3 Collaboration Processes .......................................................................................................................... 32
  3.1 Planning and allocation ............................................................................................................................ 32
  3.2 KB Development ...................................................................................................................................... 33
  4 Commonly Occurring Knowledge-Engineering Problems ............................................................................ 35
  4.1 Functions .................................................................................................................................................. 35
    4.1.1 Encoding Functions ............................................................................................................................... 35
    4.1.2 Identifying Functions in the Text .......................................................................................................... 36
  4.2 Facilitating Functions .............................................................................................................................. 40
  4.3 Qualitative Relations ............................................................................................................................... 44
  4.4 Roles ....................................................................................................................................................... 50
    4.4.1 Role Identification .................................................................................................................................. 50
    4.4.2 Role Definition and Use ......................................................................................................................... 51
  4.5 Chemistry Representation: Elements and Atoms, Reactions, Catalysts, Biomolecules, and Bonds ............... 52
  4.6 Meronymic Relations between Entities ..................................................................................................... 56
    4.6.1 Choosing a Relation .............................................................................................................................. 56
    4.6.2 Redundancy ......................................................................................................................................... 57
  4.7 Superclasses vs Subevents and Parts ........................................................................................................ 58
  4.8 Spatial Relations ...................................................................................................................................... 60
    4.8.1 Graphical Notation for Spatial Extents .................................................................................................... 60
    4.8.2 Graphical Definition of Spatial Relations ............................................................................................ 60
    4.8.3 Holes .................................................................................................................................................... 62
  4.9 Negative Information ............................................................................................................................ 63
  4.10 States ..................................................................................................................................................... 63
  4.11 Biological Classification ......................................................................................................................... 65
  4.12 Aggregates ............................................................................................................................................. 68
    4.12.1 Superclasses ....................................................................................................................................... 68
    4.12.2 Identifying Aggregates in the Text ......................................................................................................... 68
1 Introduction

This manual will help a subject matter expert (SME) formalize knowledge in a biology textbook, with AURA, in a way that it can be used for answering questions. We assume that the knowledge formulation expert (KFE) has undergone the basic AURA training and has acquired supervised hands-on experience.

The goals of this manual are to define and provide guidelines for
- Processes by which an SME encodes knowledge from the textbook, assures its quality, and reports progress (Section 2)
- Collaboration and team organization (Section 3)
- Common modeling patterns that arise in encoding knowledge (Section 4)

2 Knowledge Engineering Processes

Here, we specify the processes by which an SME encodes knowledge from the textbook, assures its quality, and reports progress. The overall flow of the subprocesses is described below, and is shown in Figure 1. In Figure 1, you can see that the process is broken into 6 phases, which can be divided into content creation determining relevance, reaching consensus, encoding planning, and encoding) and testing knowledge (key term review and question testing), all with quality assurance steps and division of labor across two teams.
Figure 1: KE Process over 6 steps with 2 teams. Time estimates are based on 2011 encoding.

2.1 Content Creation

The Content Creation process includes the first four steps of the KE Process, shown in Figure 1. SMEs encode the textbook one chapter at a time, using the following steps:

1) Determining Relevance: Select relevant sentences – The process starts by identifying a set of relevant sentences from the text. Figures are translated into sentences, and treated analogously to the text.

2) Reaching Consensus: Paraphrase sentences as universal truths (UTs) – Each relevant sentence is paraphrased as a set of UTs. A UT is a statement about exactly one entity or process.

3) Encoding Planning:
   - Create and review the taxonomy – The UTs (and the concepts to which they correspond) are used to create/review the taxonomy.
   - Plan the representation of a concept – The UTs corresponding to a given concept are used to plan its representation.

4) Encoding: Create the concept map – Finally, given the plan of a concept, this content is encoded in the system.
2.1.1 Select Relevant Sentences

Sentences are deemed relevant if they convey information about:

- Definitions of terms (often denoted around text that is bolded or italicized)
- Structure and function of entities
- Participants and steps of processes
- Concepts that are used multiple times in the textbook, or concepts that appear in the table of contents, headings, chapter reviews, and textbook glossary
- Core themes in biology:
  1. Science as a Process
  2. Evolution
  3. Energy Transfer
  4. Continuity and Change
  5. Relationship of Structure to Function
  6. Regulation
  7. Interdependence in Nature
  8. Science, Technology and Society
- Theories, laws, conjectures, hypotheses, and experiments about a core theme

Sentences are deemed irrelevant, if they convey:

- Etymological (linguistic origin of terms) information
- Metaphor and simile/analogical relationships
- Questions
- Overly detailed information
  - Some concepts in the textbook are tied to cross-curricular ideas in chemistry and physics. These ideas, although sometimes presented in detail when first introduced, are later applied only generally. Similarly, some biological ideas are presented with great detail in a very limited context, but then never revisited in such detail. In both of these cases, some or all of the detail can be omitted from encoding the idea. This should help a knowledge engineer (KE) avoid confusion about specific types or values used rarely in the text, and help appropriately present detail at the level relevant to the textbook.
  - For example, in the case of isomers, there are structural, geometric, and enantiomeric isomers, and specific types of these isomers: L, D, cis, and trans. This detail is presented in and around Figure 4.7. Isomers are presented as a case where diversity in structure can dictate diversity in function, a core theme of biology. Isomers are presented in several other places in the Campbell Biology textbook with the following detail:
    a) Phytochrome conversion between Pr and Pfr states, to regulate germination and flowering in plants (Isomer, Section 39.3)
    b) Saturated fats can occur in cis- and trans- forms (Cis- and Trans- bonds in isomers*, Section 5.3)
    c) Sugars can occur as rings or straight chains (isomers*, in and around...
Only one of these examples presents knowledge at a level below the highest taxonomic level for isomer. Enantiomer, geometric isomer, and structural isomer could be used as synonyms for isomer, along with L- and D- isomers. Cis- and trans- isomers could be created as subclasses of isomer; this was determined to be of value due to frequent discussion of cis- and trans- fats in the media, tied to physiology and nutrition. An teacher biologist can direct KE staff to encoding details such as this, as well as identifying cases where general terms are applied, as in the two examples above.

- Overly general information
  - The overview section of a chapter may have sentences that convey information in a very general way, and these sentences can be considered irrelevant.

- Examples
  - Some examples provide data that, when generalized, is redundant of an idea, or is not relevant. For instance, an idea requires a definition and the case meets the definition.
  - Specific cases that are instances of an entity (i.e., Cystic Fibrosis as an instance of an autosomal recessive condition) may be relevant as an instance of a class. A specific example of ONE TYPE of inheritance of cystic fibrosis that is conditional based on genotypes of parents, on the other hand, is not relevant; it in fact is used to illustrate a general idea or describe a concept (see prior bullet).

- Historical details
  - Founding researcher names and model organism types are relevant, but years of research, university or town names, and personal information are not relevant.

### 2.1.2 Paraphrase Sentences as Universal Truths

Given a sentence, the following steps should be performed:

1. Paraphrase the sentence as one or more universal statements
2. For each universal statement, choose the most general concept that the UT is about.

*Steps 1 and 2 above are not entirely distinct; for most cases, the universal truth directly indicates the Concept chosen, but in some cases, context indicates that the Concept would be a superclass of the Concept indicated. This means that in many cases steps 1 and 2 are done concurrently; the universal truth immediately indicates the Concept that will be encoded.*

We begin by defining universal statements more precisely, and then illustrate each of these steps with examples.
Note: The next two subsections have several examples of paraphrases. It is possible to choose different Concepts to encode a paraphrase. Please do not jump to conclusions. Wait until Section 2.1.2.3 for guidelines on how to make this decision!

2.1.2.1 Types of Universal Statements

In this section, we give an example based definition of universal truths (UTs). Each UT directly mentions a concept and optionally a context. Based on the presence of a context, we have two types of UTs:

a. Inherent UTs – Statements that are true about all instances of a concept such as an entity or an event. For example:
   a. All plant cells have a nucleus
      i. Concept = Plant cell
   b. Photosynthesis produces oxygen
      i. Concept = Photosynthesis

b. Contextual UTs – Statements that are true about all instances of a concept in a context. The context is usually a single concept but it can be a set of concepts with their own contexts. For example:
   a. The phospholipids in a membrane move laterally
      i. Concept = Phospholipid, Context = Membrane
   b. The oxygen produced during photosynthesis is used during cellular respiration
      i. Concept = Oxygen, Context = Photosynthesis
   c. Cellular respiration in eukaryotes happens in the mitochondria
      i. Concept = Cellular Respiration, Context = Eukaryote
   d. The phospholipids in the membranes of animal cells move laterally
      i. Concept = Phospholipid, Context = Membrane-Of-Animal-Cell
      ii. In this example, the context “membrane of animal cell” can be split into a concept (membrane) and a context (animal cell)
   e. During cotransport of sucrose in yeast cells, a transport protein is phosphorylated
      i. Concept = Cotransport
      ii. There are two concepts in the context – Sucrose and Yeast Cell

In the following sections, we will focus on examples on inherent UTs (such as (a)(I) and (a)(II)) or the simpler contextual UTs (such as (b)(I), (b)(II), and (b)(III)). For more complex contextual UTs, a SME should request a knowledge engineer’s assistance.

2.1.2.2 Paraphrasing Sentences as Universals

A sentence can be paraphrased as a UT in any of three ways. The first is a direct UT, where the UT and the sentence convey the same information. The second is a generalized UT, where the UT generalizes the information in the sentence. And,
finally, we have instance UTs that convey instances or examples of information in the sentence.

Note: The following are examples of UTs. It is possible to choose different most general concepts for each UT. See Section 2.1.2.3 for how to make this decision.

Note: There are two ways of classifying UTs. The first way of classification is between inherent and contextual UTs, as discussed in the previous section. The second way of classification is between direct, generalized, and instance UTs. The second classification is based on how a sentence is paraphrased as a UT. For example, a UT may be a direct UT with respect to one sentence, but it may be a generalized UT with respect to another.

We begin with examples of direct UTs:

1. Photosynthesis generates oxygen and organic molecules used by the mitochondria of eukaryotes (including plants and algae) as fuel for cellular respiration.

   a. **Photosynthesis** generates oxygen and organic molecules
      i. Inherent universal. Concept = Photosynthesis
   b. **Plants** and **algae** are eukaryotes
      i. Inherent universal. Concept = Plants and algae. Note that this example should be split into two UTs, one about plants and the other about algae.
   c. **Cellular respiration** in a eukaryotic cell happens in the mitochondria
      i. Contextual universal. Concept = Cellular respiration, Context = Eukaryotic cell
   d. **Cellular respiration** in a eukaryotic cell uses oxygen and organic molecules as fuel
      i. Contextual universal. Concept = Cellular respiration, Context = Eukaryotic cell
   e. The **oxygen** and **organic molecules** generated by photosynthesis are used by the mitochondria of eukaryotes as fuel for cellular respiration
      i. Contextual universal. Entity type = Oxygen and organic molecules, Context = Photosynthesis

2. The lateral movement of phospholipids within the membrane is rapid.
   a. UT 1 – The **lateral movement** of a phospholipid within a biomembrane is rapid
   b. UT 2 – The **phospholipids** in a biomembrane undergo rapid lateral movement
      i. Contextual universal. Concept = Phospholipid, Context = Biomembrane
3. If we place the cell in a solution that is hypotonic to the cell (hypo means “less”), water will enter the cell faster than it leaves, and the cell will swell and lyse (burst) like an overfilled water balloon.
   a. A cell in a hypotonic solution will gain water faster than it loses it
      i. Contextual universal. Concept = Cell, Context = Hypotonic-Solution
   b. A cell in a hypotonic solution will swell and lyse (burst) like an overfilled water balloon
      i. Contextual universal. Concept = Cell, Context = Hypotonic-Solution

The first sentence is paraphrased as several distinct UTs. For the second case, there are two choices of universal paraphrases which may lead to different encodings. Finally, the third example illustrates how conditional statements can be paraphrased as universals. See Section 2.1.2.3 on how to choose the most general concept for such UTs and/or whether or not they should be encoded.

The second type of paraphrase is where we generalize the information conveyed by the sentence:
   i. The most abundant lipids in most membranes are phospholipids
      a. All membranes have more phospholipids than other lipids
         i. Inherent universal. Concept = Membrane
      ii. One way adenosine triphosphate (ATP) can power active transport is by transferring its terminal phosphate group directly to the transport protein.
         a. Active transport obtains power from ATP transferring its terminal phosphate group to the transport protein
         i. Inherent universal. Concept = Active-Transport

Observe that in the first case we generalized a sentence about “most” membranes to “all” membranes. In the second case, we generalized a sentence about one way that ATP powers active transport to a statement about all occurrences of active transport. A teacher biologist should decide if a generalized paraphrase is appropriate.

The final method of paraphrasing is to instantiate or skolemize a sentence:
   1. Some prokaryotes even contain a row of tiny magnetic particles that allow the cells to orient in Earth's magnetic field.
      – Magnetotactic bacteria contain a row of tiny magnetic particles that allow cells to orient in the Earth’s magnetic field
         * Inherent universal. Concept = Magnetotactic-Bacteria

Here, we have chosen “magnetotactic bacteria” as an instance, example, or skolem for “some prokaryotes”. Only those examples that are mentioned in the textbook are used. This method of paraphrasing can be used for sentences that (a) contain words such as “some”, “may”, “possibly”, “can”, and (b) cannot be paraphrased as universals by generalizing.
2.1.2.3 Choosing a Concept to Encode a Paraphrase

We choose exactly one most general concept (MGC) for each UT. The information in the UT will be associated with this concept in the knowledge base. For inherent universals, the MGC should be the concept that is directly mentioned in the sentence or one of its superclasses. For example:

i. All **plant cells** have a nucleus
   a. Concept = **Plant-Cell**
   b. The MGC should be Plant-Cell or one of its superclasses
   c. In this case, MGC = Eukaryotic-Cell. The information in this UT will be associated with the concept Eukaryotic-Cell and since Plant-Cell is one of its subclasses, we will infer that all plant cells have a nucleus.

ii. **Photosynthesis** generates oxygen and organic molecules
    a. The MGC should be photosynthesis or one of its superclasses

iii. **Plants** and **algae** are eukaryotes
    a. Plant or one of its superclasses should have eukaryote as its immediate superclass
    b. Algae or one of its superclasses should have eukaryote as its immediate superclass

The MGC for an inherent UT should never be encoded on the subclass of the entity/event that the paraphrase is about! If there is a need to do that, then the paraphrase is incorrect.

For contextual universals, there are two choices for the MGC:

i. The contextual entity/event (or one of its superclasses)

ii. A concept that reifies the entity/event with its context (or one of its superclasses)

If an SRI Biologist deems that reification is appropriate (choice ii), then it should be used. Otherwise, use the context choice, as appropriate. Note that in some cases neither choice may be appropriate, and such sentences should be labeled as a knowledge representation (KR) issue.

Consider the following examples from the previous section:

i. **Cellular respiration** in a **eukaryotic cell** happens at the mitochondria
   b. MGC from context – Eukaryotic-Cell or one of its superclasses
   c. MGC by reification – Cellular-Respiration-In-A-Eukaryotic-Cell or one of its superclasses.
   d. In this case, teacher biologists prefer reified MGC. Generally speaking when the directly mentioned concept is an Event and the context is an Entity, the reified choice is usually preferred.

ii. **Cellular respiration** in a **eukaryotic cell** uses oxygen as a fuel
   a. Same as the previous paraphrase
iii. The **oxygen** and **organic molecules** generated by **photosynthesis** are used by the mitochondria of eukaryotes as fuel for cellular respiration
   b. MGC from Context – Photosynthesis or one of its superclasses
   c. MGC by Reification – Oxygen-Generated-By-Photosynthesis or one of its superclasses. And, Organic-Molecule-Generated-By-Photosynthesis or one of its superclasses.
   d. In this case, we prefer the MGC based on the context Concept, because the reification leads to unnatural concepts.

iv. **UT 1** – The **lateral movement** of a **phospholipid** within a **biomembrane** is rapid
   a. Contextual universal. Event type = **Lateral movement** Context = **Phospholipids-Within-A-biomembrane**
   b. In this case there are three choices of MGC:
      i. MGC from context 1 – Biomembrane or one of its superclasses
      ii. MGC from context 2 – Phospholipid-Within-A-Biomembrane or one of its superclasses.
      iii. MGC by reification – Lateral-Movement-Of-Phospholipid-Within-A-Biomembrane or one of its superclasses.
   c. Since the concept is an event and the context is an entity, we are forced to choose the reified MGC, i.e., Lateral-Movement-Of-Phospholipids-Within-A-Biomembrane. However, this is a very unnatural concept, and so the SME should either write a different UT or mark the sentence as a KR issue.

v. **UT 2** – The **phospholipids** in a **biomembrane** undergo rapid lateral movement
   a. Contextual universal. Entity type = **Phospholipid**, Context = **Biomembrane**
   b. MGC from context – Biomembrane or one of its superclasses
   c. MGC by reification– Phospholipid-Within-A-Biomembrane or one of its superclasses
   d. In this case, the biology teachers preferred the MGC from the context.

vi. A **cell** in a **hypotonic solution** will swell and lyse (burst) like an overfilled water balloon
   a. Contextual universal. Entity type = **Cell**, Context = **Hypotonic-Solution**
   b. MGC from context – Hypotonic Solution
   c. MGC by reification – Cell-In-A-Hypotonic-Solution
   d. In this case, the MGC based on context is not appropriate, because hypotonic solutions do not necessarily involve cells, and the MGC based on reification is unnatural. So, it should be marked as a KR issue. [Editor's note: Currently, the knowledge base (KB) uses the MGC based on reification, but it is likely that this class will be removed and the sentence left as a KR issue.]
2.1.3 Create and Review the Taxonomy

The set of paraphrases determines the set of concepts to be created. A top-down taxonomy (also known as a class hierarchy) organizes the concepts from general (Cell) to specific (Plant-Cell). A taxonomy is created by assigning a direct superclass to each concept, and writing a definition for the concept.

In the following sections, we discuss guidelines for assigning a direct superclass, naming conventions, writing concept definitions, and finally, reviewing the taxonomy.

2.1.3.1 Guidelines for Assigning a Direct Superclass

Use the following rules when assigning a direct superclass:

a. Make sure that the class inherits all the properties of the superclass, and choose the most specific superclass. For example, Cell and Eukaryotic-Cell are both candidate superclasses for Plant-Cell. Since Eukaryotic-Cell is more specific (it is a subclass of Cell), it is chosen as the direct superclass of Plant-Cell. Use the following tests to decide whether X is a superclass of Y:
   a. Is Y a kind of X? Is a Protein a kind of Molecule?
   b. Is Y a type of X? Is a Nucleus a type of Organelle?
   c. Is every object that is an instance of Y also an instance of X? Is every occurrence of Phosphorylation also an occurrence of a Chemical-Reaction?

b. Check that the concept definition is consistent with the definition of the superclass.

c. AURA allows multiple direct superclasses. For example, Protein can have Macromolecule and Amphipathic-Entity as its direct superclasses. Multiple inheritance should be used with care. The SME should ensure that the class does inherit all the information in both superclasses in a compatible way. See Section 4.7 for further guidelines.

d. A particularly difficult case is whether to choose an entity or a role as a superclass for a concept. See Section 4.4 for guidelines.

2.1.3.2 Naming Conventions

Defining naming conventions for concepts in a KB and then strictly adhering to these conventions makes it easier to understand the KB and also helps avoid some common modeling mistakes. The following guidelines must be followed for naming new concepts for an AURA Biology KB.
1. **Capitalization.** Use consistent “Title-Case” capitalization for concept names. Every word after the hyphen is to be capitalized. For example, if you want to define a concept for eukaryotic cell, name it Eukaryotic-Cell.

2. **Multi-word names.** If there is a concept whose name involves two or more words, follow uniform naming conventions. For example, the second word of the concept should be separated with a hyphen and start with a capital letter. For example, ‘Startcodon’ and ‘Ribosomalsubunit’ should be named ‘Start-Codon’ and ‘Ribosomal-Subunit’, respectively.

3. **Sets.** A class name represents a collection of objects. For example, a class Cell actually represents all the cells. *There is no need to add obvious plurals* as synonyms to a concept. *Non-standard plurals should be added* in the word-to-concept mapping. For example, the plural of “nucleus” is “nuclei”, and in such cases, the concept should be named after the singular form, but the plural form should be added to the word-to-concept mapping.

4. **Names.** Do not add strings such as “class” to concept names. For example, Eukaryotic-Cell, but not Eukaryotic-Cell-Class.

5. **Abbreviations.** Do not use abbreviations in concept names (that is, use Eukaryotic-Cell, and not Euk-Cell).

6. **Prepositions** such as of, in, and so forth can be included in the concept names, when they add clarity or make the name easier to understand.

7. **Biological names.** Use the name that is commonly used in the text, whenever possible, or the natural name. Thus, associated properties may come either before or after the entity name. Thus, ‘Smooth-Endoplasmic-Reticulum’ is more natural than ‘Endoplasmic-Reticulum-Smooth’.

8. **Distinct names.** Avoid giving the same name or very similar sounding names to two biological concepts. For example, there should be only a single concept for ‘Homologous-Chromosome’ and ‘Homologous-Chromosomes’.

9. **Name consistency.** Names of direct subclasses of a class should either all include or all not include the name of the superclass. For example, if we are creating two subclasses of Organism to represent eukaryotic and prokaryotic organisms, the two subclass names should be either Eukaryotic-Organism and Prokaryotic-Organism or Eukaryote and Prokaryote, but not Eukaryotic-Organism and Prokaryote. For another example, consider naming the sibling concepts representing DNA replication in eukaryotes and prokaryotes. We should either use the names DNA-Replication-In-Eukaryote and DNA-Replication-In-Prokaryote, or Eukaryotic-DNA-Replication and Prokaryotic-DNA-Replication. The choice can be made after consultation with SRI.

10. **None of the above.** When there is no specified rule or advice above, then SMEs should discuss the possible names among themselves and with SRI to come up with a naming convention for the cases in question.
**Exceptions** to the above guidelines: Some biological entities are expressed in either capital letters (e.g., acronyms) such as ‘TATA Box’, or start with a digit such as ‘5-Prime-End’ and ‘5-Prime-Cap’. They should be named identically to their naturally occurring name. The usage found in the text should be used in the name. If the abbreviated name is more common, then this is what should be used in the concept map. For example, names such as NADP, ATP, and CO2 could be used directly. The expanded name must be included in the synonym list.

### 2.1.3.3 Concept Definition

Every concept should have a definition. There are two cases to consider:

i. If the concept has an entry in the glossary of the textbook, the definition in the glossary should be used.

ii. If the concept has no glossary entry, the definition should be in the style of a glossary entry. This can be determined by looking at glossary entries for superclasses or siblings in the taxonomy. For example, to define Cellular-Respiration-In-Eukaryote, the biologist may consult the glossary entry for Cellular-Respiration and then, in a similar level of detail, describe what is unique about Cellular-Respiration-In-Eukaryote.

A definition should present basic knowledge about a concept within the scope of the textbook. All definitions are three sentences or less. If a definition is not provided in the book, then it should be authored by an SRI teacher biologist (without plagiarism from other published sources). If plagiarized content (i.e., content from Wikipedia) is present in either the comments or description field, this should be deleted.

**Where to put the concept definition:**

The Comments field has been populated by extracting glossary entries from the Campbell Biology textbook. It should never be edited. Whenever this field is populated, the software will prefer this value over the value in the Description field. As a matter of policy, we never update the values in the Comments field as they were obtained from the textbook over which we have no control. We only update the description field.

The description field needs to be given only for those concepts for which no entry in the comments field exists. In other words, these are concepts in the knowledge base with no corresponding entries in the glossary of the textbook. This also means that if the users are not satisfied with the entry in the Comments field, no action is possible, because we never edit this field and this field is preferred over the value in the description field. We have made this decision to bound the level of
effort/refinement invested in authoring the documentation strings. Finally, as noted above, the description field will be authored by the teacher biologists at SRI.

### 2.1.3.4 Reviewing the Taxonomy

This description of some desirable and undesirable properties of a taxonomy will help an experienced SME review it. In addition to these properties, the reviewer should verify that a concept has an appropriate definition,

**Cycles:** The class hierarchy should be acyclic; that is, there should be no cycles. We say that there is a cycle when some class A has a subclass B, and B is also a superclass of A. Creating such a cycle amounts to declaring that the classes A and B are equivalent: all instances of A are instances of B and all instances of B are also instances of A. AURA checks for cycles and prevents their creation. In the past, there have been cases where cycles were created in the hierarchy in spite of such checks. Therefore, it is important for SMEs to be aware of this representation requirement for debugging.

**Siblings:** Siblings in the hierarchy are classes that are direct subclasses of the same class. All the siblings in the hierarchy (except for those at the root) must be at the same level of generality. For example, Vacuole and Food-Vacuole should not be subclasses of the same class (say, Organelle). Vacuole is a more general concept than Food-Vacuole. Siblings should represent concepts that are at the same level, in the same way that same-level sections in a book are at the same level of generality. In that sense, requirements for a class hierarchy are similar to the requirements for a book outline. However, the concepts at the root of the hierarchy (which are often represented as direct subclasses of some very general class, such as Thing) represent major divisions of the domain and do not have to be similar concepts.

There are no hard rules for the number of direct subclasses that a class should have. However, many well-structured taxonomies have two to a dozen direct subclasses. Therefore, we have the following two guidelines:

- *If a class has only one direct subclass, there may be a modeling problem or the taxonomy is not complete. After checking for modeling and taxonomy issues and determining that there are no siblings for the subclass, the KFE should make the subclass name simply be a synonym of the superclass.*

- *If there are more than a dozen subclasses for a given class, additional intermediate categories may be necessary.*

The first guideline is similar to a typesetting rule that bulleted lists should never have only one bullet. For example, most red Burgundy wines are Côtes d'Or wines. Suppose we wanted to represent only this type of Burgundy wine. We could create a
class Red Burgundy and then a single subclass Côtes d’Or. In this case, red Burgundy and Côtes d’Or wines are essentially equivalent (all red Burgundy wines are Côtes d’Or wines and all Côtes d’Or wines are red Burgundy wines), creating the Côtes d’Or class is not necessary and does not add any new information. If we included Côtes Chalonnaise wines, which are Burgundy wines from South of Côtes d’Or, then we will create two subclasses of Red Burgundy: Côtes d’Or and Côtes Chalonnaise.

The second guideline is illustrated by the following example. There are numerous subclasses of enzymes such as Carbonic Anhydrase, Lactase, Dehydrogenase Oxidases, DNA Polymerase, Transaminase Kinases, Decarboxylases, Aldolases, Cellulases, Amylases, Xylanases, Lipases, Lactases, Methylase, Creatine Kinase, and Alanine Aminotransferase. Instead of creating so many subclasses of the ‘enzyme’ class, an additional level of classification can be created based on the manner in which these enzymes get formed. Thus, the various enzymes can be grouped under six major subclasses – namely, Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases.

However, if no natural classes exist to group concepts in a long list of siblings, there is no need to create artificial classes – just leave all the classes. After all, the ontology is a reflection of the real world, and if no categorization exists in the real world, then the ontology should reflect that.

2.1.4 Concept Representation Planning

Once a reviewed taxonomy is available, we plan the representation for the most important concepts. The important concepts are determined by consultation with SRI. The main goal during the planning phase is to identify and make important choices before encoding begins. Specifically, we wish to check the UTs so that we avoid (a) the creation of overly complex Concepts and (b) missing crucial information.

2.1.4.1 What is a plan?

A plan is a sketch of how the main ideas are conveyed on the Concept of a concept. The plan is made by an experienced SME, designated as the planner, and difficult issues are resolved by discussion with all the SMEs and KEs. The plan may be made using paper drawings, AURA, or any other tool.

As we mentioned in the previous paragraph, a plan is a sketch of the main ideas. The main ideas in an entity Concept may include information about structure and function. For example, a phospholipid bilayer is a substructure of a biological membrane, and one of its functions is to transport molecules into and out of a cell or organelle. Similarly, the main ideas in an event Concept may include information
about participants and steps. For example, glucose is a participant (raw-material) in cellular respiration, and glycolysis is a step.

### 2.1.4.2 How to create a plan?

To determine the main ideas for a concept, the planner should select the UTs for that concept, its superclasses, and related concepts. The plan describes the information that is local to the concept, and relevant information that is inherited or placed on other Concepts. For example, we may decide to place certain proteins as parts of a membrane, and encode other information in subclasses of membrane proteins. Note that the choice of Concepts should be motivated by the UTs! If a UT leads to an unsatisfactory encoding, it should be rewritten.

Creating a detailed plan of all information in a Concept is impractical. The planners should focus on the main ideas, as we have illustrated above. Since the plans are sketches or designs, there may be unforeseen issues during encoding and testing. These issues can be corrected during those phases of the process. The main goal during planning is to check whether the Concepts are at the right level of detail, given the UTs. We should try to avoid (a) the creation of overly complex Concepts, if possible, and (b) missing crucial information.

### 2.1.5 Concept Map Creation

Concept maps should be created top-down. That is, before we create a Concept, we should create the Concepts for all its superclasses. Concepts are created by an SME, designated the encoder. The encoder should use the following steps:

- Select all the UTs associated with the concept.
- Consult the plans, if any, for the concept.
- Write action items for each UT. The action items should include all the encoding that will be carried out based on that UT.
- Create or edit the Concept in AURA as indicated by the action items.

SMEs create Concepts for entities, events, and roles. Each type of concept has important slots associated with it. For example, entities have associated meronymic relations (has-part, has-region, and material), spatial relations (is-between, is-inside, is-outside), properties (size, area, diameter, volume), and relations associated with its function (has-function and facilitates). To use slots, the SMEs should be familiar with (a) the definitions of the slot in the description pane in AURA, and (b) the guidelines for specific modeling patterns in Section 4.

During Concept creation, we should take care to encode:

- **Disjoint classes**: AURA allows you to specify explicitly that several classes are disjoint. AURA checks for disjoinededness and prevents creation of a class that is a subclass of two disjoint classes. Classes are disjoint if they cannot have any instances
in common. For example, tRNA, rRNA, and mRNA in our ontology are disjoint and they should be encoded specifically as disjoint sets. One way to identify classes as being disjoint is to determine whether they are used exclusively of each other in classification. For instance, eukaryotic and prokaryotic cells are disjoint because if a biologist knows a cell is not a prokaryotic cell, then it must be a eukaryotic cell. If prokaryotic cell (aka prokaryote) and eukaryotic cell are made disjoint, then AURA can reason the same way. Similarly, Organism has subclasses Unicellular-Organism (this Organism is also a Cell) and Multicellular-Organism (this Organism has many cell parts), which can be considered disjoint.

Note that it is not appropriate to always think of subclasses in a taxonomy as being disjoint from each other, however. For instance, Living-Entity has subclasses Cell and Organism. As described above, Unicellular-Organism is one type of Organism that can be used to show that sibling classes are not disjoint. In this case, both classes, Cell and Organism are superclasses to the same Unicellular-Organism.

- **Sufficient properties (Triggers):** A set of *sufficient properties* of a class is a set of properties that, if satisfied by an instance, are sufficient for it to be a member of that class. For example, if a Cell has a Nucleus, that fact is sufficient for it to be a member of the class Eukaryotic Cell. In AURA, the sufficient properties are encoded by adding triggers to concepts. If several features are added to a single trigger, then all of these features must be true for that trigger to be used in reasoning. If there are two separate triggers for a concept, however, these are disjunctive, meaning that only one has to be satisfied in order for the trigger to be used in reasoning. Definitions of concepts are useful for determining triggers; they indicate how a concept differs from its superclasses and subclasses, and indicate how it might be classified.

### 2.2 Methodology for Testing Knowledge

The primary method for testing knowledge in AURA is by posing questions and comparing the AURA response with the correct answer. In KB development, testing happens at two levels: during knowledge encoding and after knowledge encoding.

#### 2.2.1 Testing During Knowledge Encoding

During knowledge encoding, the SMEs encode concepts belonging to various depths of knowledge. The concepts and their corresponding testing procedure are described here:

- **Small concepts:** These concepts are simple and have mainly ‘has-part’ relations or an event representation. The depth of Concept is not more than two levels. In these concepts, start the debugging process after the concept is fully encoded. As an example, for the concept *Codon*, we can check for errors by asking questions related to it after it is fully encoded.
• Medium concepts: These concepts have a larger number of components and multiple events. For example, “Genetic variation” has three subevents: independent-assortment, crossing-over, and random-fertilization. These subevents themselves have few entities in their Concepts. For such concepts, SMEs should check the encoding at two stages – first, after completing almost half of the encoding and then again after completing the encoding. Both times, questions can be asked to test the encoding.

• Large concepts: These concepts are complex and involve several medium-size (or large) concepts. The connections among the entities and events are also numerous. For example: meiosis involves several subevents, such as nuclear division, cytokinesis, telophase, anaphase, metaphase, and prophase. Testing has to be done in more detail. The SMEs should encode one subevent and check that subevent with two or three questions. Then, when the next event is encoded, again multiple questions for that event should be tested. Further, the first subevent’s questions should be again retested to make sure they still work. Follow these detailed steps:

1. Decide the stages at which you will test/debug the concept. For example, if meiosis has five important subevents, debug at the first level to check if the system can name the subevents correctly. Then add the concept map for the first subevent and ask a question about it. Add the remaining subevents one by one and ask a question after each subevent has been added. In this case, questions will need to be done six times (once for the first level, and five times for the five separate subevents).

2. Start with very simple, direct questions, such as the following:

   o What are the parts of X?
   o What are the subevents of X?
   o What is the agent of X?
   o What is the result of X?

3. When a question is asked, first check the graphical feedback from AURA to verify whether the question was correctly interpreted by the QA system. Any unbroken links or something out of place in the question graph should raise concern about the encoding or style of the question. The textual paraphrase present on the right-hand screen should also be checked for verification.

4. After successful testing on simple questions, test complex questions. These questions would require the system to go deeper into the concept maps and search for specific information. We sometimes lead AURA to the exact point in the concept map before asking the main question. After AURA responds in a satisfactory manner, ask a more complex but shorter question that just provides the context in which it is being asked. Two things need to be checked: (a) the graphical interpretation of the question
should correctly reflect the connection between various entities in the KB, and (b) AURA should ‘infer’ the right answer.

5. Tests of adherence to modeling guidelines have been automated, and this automated testing suite will continue to be revised as guidelines are improved and old errors in the KB are fixed. After saving encoding, the encoder runs these tests by the following: in the AURA editor pane, from the View menu, select Test Encoding Quality. The Encoding Quality Results for the Concept currently being viewed in the AURA editor is generated, which includes the test run and the corresponding result. “No encoding issues found” indicates that the algorithms tested were not violated. If a violation is found, the encoder should review all of these and determine if additional changes should be made to that Concept, or if the violations are acceptable. The encoder can subsequently modify the encoding or file/edit a JIRA issue if this is a known issue that needs others’ input. The automated test suite checks for things such as the following, with reference to guidelines in this KE Manual:

- Appropriate use of relations based on previously indicated domains and ranges (e.g. has-part, has-region, and possesses) – See section 4.6.1
- Appropriate number of slots to or from relations (e.g. exactly two relations from is-between) – Guidelines not in this KE Manual
- Violations such as poor unification – Guidelines not in this KE Manual
- Bad filenames – See section 2.1.3.2
- Bad superclasses (e.g. Thing) – See section 2.1.3.1 and 2.1.3.4
- Roles with non-permitted relations – See section 4.4

You should rate the quality of the 'biology' related output of a question. The various parameters that would determine the quality of the biology output are still evolving. As of now, the following parameters should be used:

a. Check for extraneous information – does the result string contain extra information?

b. Make a note if the final answer is still vague.

c. Check for explanation details – does the explanation provided below the answer have sufficient details?

d. Check for ambiguous information in the answer. Sometimes AURA picks up results from related concepts when it should have given only one correct answer.
e. Assess whether the answer quality is purely an issue of KB content. If yes, then suitably modify the KB content to get the right output. The following steps can be used to understand whether it is a KB content issue:

- Is the knowledge that you are expecting in the answer explicitly represented in the KB?
- Does the KB contain more or less information than what is expected in the answer? If the KB contains less information, can the missing information be derived (e.g., by inheriting from superclasses)?
- Was the question understood correctly?

2.2.2 Testing After Knowledge Encoding

Once the knowledge base is finalized, it is tested through a minimal test suite. In the previous iterations of the KE process we have also used reference test suites and the novel question test suites. We describe all three test suites, but at present, we are not using the reference question suite and the novel question suites.

1. Minimal Test Suite: This includes all key terms from an encoded chapter within “What is…” questions, and at least one question of each deeper reasoning type that has been implemented, as long as this content occurs in the chapter being encoded. To write the deeper reasoning questions, the SMEs use templates associated with that question. SMEs are trained on the current answer presentation for these questions, so known JIRA issues can be appended as needed, and new JIRA issues are only filed for new issues in answer presentation. Once biologists agree that a deeper reasoning question is “passing,” a screen shot of it is taken, so it can be archived for future comparisons. As AURA matures and answers are regenerated, review of old questions and answers are an important aspect of the quality assurance process. Templates include:

- Similarity and difference questions between:
  - Two entities not related
  - Two subevents for the same event
  - Two forms of entities (e.g. isomers)
  - Superclass and subclass
  - Two similar entities
  - Functions of two entities
- Relationship questions between:
  - Two entities – a whole and a part
  - Two events that have a next-event relation between them
  - Two events that have a causal relation between them
  - Event and entity not directly connected to each other
  - Two properties that have a qualitative relationship
  - An event and a role
• Process Interruption (This question template is not yet implemented)
  o If A is removed from B, what events will be affected?
  o If A is removed from B, will event C be affected?
• Similarity Reasoning
  o X is to Y as Z is to what?
• Structure / function relationship
  o What structure of X facilitates the function of X?
  o What structure of X facilitates function Y?

2. Reference Test Suite: These questions are more complex than minimal test suite questions. Their testing involves checking of more relations and concepts. The queries framed for these questions are longer, and more attempts and twisting of the CPL queries are required to lead to an answer.

3. Novel Questions: These questions are indirect and test the depth and breadth of the knowledge base. The SMEs may have to change the wording of the queries several times to get them to answer correctly.

Generating these test suites is also an important task and is done in collaboration with the teacher Biologists. The SMEs follow templates to generate questions in the minimal test suite, after which questions are reviewed and sometimes supplemented or edited by teacher Biologists.

When testing is done by the SMEs who encoded the knowledge, a level of subjectivity and leniency creeps in. Hence, the SMEs should follow a strict procedure for testing and should assign 'correctness' to any question only after determining that the answer is complete, correct, and appropriate for an early college/advanced high school audience. This is quality assured by teacher Biologists and can be followed up by review from external evaluators, such as educators or students not involved in encoding.

Check whether AURA is giving ambiguous/multiple answers: There may be ambiguous or multiple answers being generated that contain the correct answers within. In such cases, SMEs should not assume the answer to be correct. The observations and issues should be sent to KE/KC, and a JIRA issue should be filed.

For all Novel Questions:

1. Multiple queries must answer correctly: Several queries can test the same piece of encoded knowledge. Some of these queries may not result in the same correct answer as other queries. Hence, the SMEs should never ask just one query. At least three types of different queries should be framed to check each test suite question. Evaluators should assign correctness to a question only after multiple questions are found to be correct. Any new and odd results should be immediately sent to the knowledge engineer / knowledge component (KE/KC), and a JIRA issue should be filed.
2. **Ask your peer SMEs to check some questions**: Your peer SME can come up with a very different computer-processable language (CPL) query to test a question that you have found to be correct. Do this for some randomly selected questions. Such checks further reinforce correctness.

### 2.3 Quality Assurance

The goal of the quality assurance process is to ensure a high-quality KB and answer outputs that a consumer will accept. Three specific quality assurance processes should be followed, as shown in the **Figure 2**.

![Quality Assurance Process](image)

**Figure 2: Quality assurance process**

#### 2.3.1 Concept Map Review by Biologists

The teacher biologists review the concepts developed by the SMEs and give feedback on the accuracy of the representation. The biologists should follow these guidelines:

1. **Check the class hierarchy**: Are the superclasses and subclasses correct for each element? Check whether entities require additional superclasses or subclasses.
   - Are classes identified as disjoint as appropriate?
2. **Check whether the concept map adequately represents the biological entity or event**: There are several cases where the encoding captures the knowledge in bits and pieces. Hence, one should verify the following:
   - The concept map as a whole should convey biological meaning.
   - Are the relations and properties appropriate and sufficient for the knowledge present in the textbook? Are these used in triggers?
   - Should more information be encoded for the entity?
3. **Check for redundant knowledge**: Unnecessarily encoding the same knowledge in several concepts is not uncommon. Teacher biologists have the knowledge to assess this. Note that in Section 4.6.2 there are guidelines on allowed and disallowed redundancy with meronymic relations; relations between larger and
smaller concepts should be specified at the level of the larger concept, if this is always true. For instance, a eukaryotic cell always has a nucleus, and a nucleus is always part of a eukaryotic cell. The former should be encoded, but the latter should not.

4. **Check for synonyms:** The teacher biologists are aware of the common words used to denote a particular entity or event. These words must be included as synonyms for the corresponding entity.

5. **Check the description string:** The teacher biologists should check that the description string contains relevant and adequate biology knowledge. Extraneous information should be deleted from the description string. All the relevant knowledge in the description string must be encoded in the concept map, and vice versa – all knowledge present in the concept map must be in the description string.

**2.3.2 Answer Output Review by Biologists**

SRI Biologists should review the answer output using the following guidelines:

1. **Check for the correct answer:** The system should generate the correct answer. There should be no ambiguity in the answer generated. Further, the answer generated should be backed up by reasonable explanations.

2. **Check the way the answer is displayed:** AURA should display the answer in a way that is easily understandable. Any third person should easily decipher the answer.

3. **Check for compatibility with AURA’s textual interpretation and QFC Map Editor:** The answer generated should be in line with AURA’s textual interpretation.

**2.3.3 Knowledge Representation Quality Analysis**

The KR quality analysis assesses the quality of the knowledge representation from a different perspective – that of a knowledge engineer. There is some overlap with the analysis performed by the SRI Teacher-Biologists, but we anticipate that a KE will identify different (perhaps non-biological) issues.

The steps in this section, done by a knowledge engineer, should focus on obvious mistakes, not involving detailed biological knowledge. The KE is not expected to verify the biological correctness of the representation. The KR quality analysis follows these steps:

a. **Check if the documentation string is defined properly.** The documentation string should be an overview of the knowledge encoded in the concept map and should mention the basic information about the entity. It should also reflect any information that was not possible to include in the concept map. AURA sometimes uses the documentation string as part of the answer/explanation.
b. **Check if the concept is linked to its definition in the document.** The KB concept should link to the material in the text CD. One can check the link by clicking on the concept in the editor pane and looking for the link to the specific section in the documentation pane. When found, the corresponding text in the document pane is highlighted in yellow and bold.

c. **Check for the appropriate superclass.** Each concept in the KB has a superclass and it is important to assign it appropriately. Concepts may either be assigned too general (e.g., Thing) or too specific superclasses. If there is more than one superclass, check to make sure that they are not redundant. For example, if Eukaryotic Cell has been assigned superclasses Cell and Living Entity, the superclass of Living Entity is redundant because it will be inherited from Cell.

d. **Check whether the necessary subclasses are defined.** The number of subclasses of an entity should be more than 1 and less than 12. If there is only one superclass, then we should delete that superclass and recheck the taxonomy. If an entity has more than 12 superclasses, we should create more intermediate subclasses and reorganize the taxonomy.

e. **Check whether the key properties of the event or object as defined in the text appear in the graph.** By key properties, we mean any piece of information that is important in that it might be used in answering a question. The description string and the encoded knowledge should mirror each other (fuzzily).

f. **Check if all the relations used in the concept have been used correctly.** First, the relations should reflect what is present in the information string. Second, check the correctness of the use of each relation by clicking on the specific “relation” and viewing its documentation in the left bottom/middle corner of the edit panel. Verify that the current use of the relation conforms to AURA’s rules and axioms. Finally, check the direction of the relation; using the wrong direction is a common error.

g. **Check if the constraints that are specified in the text are correctly represented.** In addition, check that the constraints are present in the description string.

h. **Check if the defining or trigger properties of a concept are represented in the graph.** First, determine if a trigger is needed. If so, check that the trigger exists and is correct – i.e., that the right nodes are connected for the trigger.

i. **Check if any links are missing:** Check for cases where nodes are unconnected or need unification.

j. **Check if there is more information represented than appears in the text.** If so, first check if the information is relevant in the context of the scope of the textbook. If it is relevant, add the extra information in the documentation string. If the information encoded is not relevant, then the SMEs can remove that encoding from the concept.

k. **Check to see if the nodes in the graph are at the right level of specificity – not too general and not too specific.** For example, if an event involves bound-ribosome,
then the encoding should show the concept 'bound-ribosome' rather than 'ribosome'.

Here we provide an example of the application of this KR quality analysis for the incorrect encoding of the concept *Anaphase* shown in Figure 3.

![Figure 3: An incorrect and incomplete encoding for ‘Anaphase’ to be used for a KR analysis.](image)

The summary of the analysis results is as follows:

i. **Missing Knowledge:** cell elongation values should be mentioned. Microtubule shortening values should be mentioned. Should there be a specific value of the ‘tangible entity’? Microtubule Shortening and Cell Elongation are incomplete in their linking of the from-value and to-value relations to the property values of their respective objects. The participants of the subevents of Anaphase are not related to each other.

ii. **Incorrect Knowledge:** Two 'moves' are connected as the first-sub event in 'move-apart’ event. The base and the object of the first 'attach' event are the same. Should have specific values for the 'value' property of the 'microtubule shortening'. There is no conceptual link between the 'detach' and 'shortening' events. Several unifications of concepts such as 'cell' are possible. *Detach* is the wrong concept to use to describe the first step, and...
*sister chromatics* do not make sense as the base and the object. *Growth* should probably be *Grow*.

iii. **Note:** There are many incomplete relations, and this concept needs to be entirely reworked. The knowledge engineer is not sure of the importance of relating participants from the various subevents; perhaps a biologist needs to be consulted. While it is technically missing, it needs to be included only if it makes a difference for QA.

### 2.4 Reporting Progress

The goal of these processes is to measure the progress and the effectiveness of the encoding, and to track with the progress of the content creation. We describe below four processes for reporting progress.

#### 2.4.1 Coverage Analysis

The objective is to ensure that all the knowledge present in the text relevant to the Biology syllabus has been captured, or that we understand the reasons why it is not captured.

The SMEs will perform this analysis by going through the text sentence by sentence and analyzing to what extent the knowledge has been captured in a concept map. It is not necessary for the concept map to capture the text word by word. Therefore, the process of assessing coverage is subjective and we expect that SMEs will get better at it as they gain experience.

The coverage analysis will also be undertaken for diagrams. The procedure for coverage analysis for diagrams (also known as Diagram Analysis) will be as follows:

- a. Write down a series of sentences in English that capture the Biology relevant knowledge conveyed by a diagram, inclusive of the diagram’s title and caption, and other text in the diagram.
- b. For each sentence, determine if it is repeated in the main body of the text, and specify the associated textbook sentence.
- c. If not, indicate that it is “not covered”.

Following this analysis, items a, b, and c can be used in sentence-based encoding for a chapter.

The coverage analysis will be accompanied by summary statistics. Specifically, the following statistics should be reported for the sentence-by-sentence analysis:

1. Number of All sentences
2. Number of Critical sentences
3. Number of Critical sentences-CK
4. Number of Non-critical sentences
5. Number of Non-critical sentences
6. Number of Non-critical sentences-GK
7. Number of Redundant sentences
8. Number of Encoded sentences
9. Number of Semi-encoded sentences
10. Number of sentences not encoded
11. Number of sentences not encoded due to AURA limitations
12. Number of sentences not encoded due to a KR issue

As the process includes planning during which each sentence is rewritten as one or more UT, statistics regarding encoding can be reported at this level of detail and combined with the statistics at the sentence level, as listed above. At this UT level, statistics include
1. Number of all UTs
2. Number of UTs that have action items made
3. Number of UTs that have action items not made
4. Number of UTs for which action items have been encoded during this pass through the chapter
5. Number of UTs that were already encoded prior to this pass of the chapter (Already Present)
6. Number of UTs that have action items that have not yet been encoded
The example analysis in Table 1 highlights the dynamic process of planning and encoding at the sentence level:

<table>
<thead>
<tr>
<th>Relevancy</th>
<th>Status</th>
<th>Sentence</th>
<th>Universal Truth</th>
<th>Concept Map</th>
<th>Encoder status</th>
<th>EVS Biologist Comments</th>
<th>Biologist UT/CM/Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>Encoding Complete</td>
<td>Heterotrophs obtain their organic material by the second major mode of nutrition.</td>
<td>Heterotrophs receive organic material by mode of heterotrophic nutrition.</td>
<td>Heterotroph</td>
<td>1/31/2012 Aasim: Debbie/Nikhil we should create a new class Heterotrophic Nutrition or we should encode in the Concept of Heterotroph a nutrition event?</td>
<td>1/11/2012 Shivani: Reviewed</td>
<td></td>
</tr>
<tr>
<td>Relevant</td>
<td>KR issue</td>
<td>Unable to make their own food, they live on compounds produced by other organisms (hetero- means “other”).</td>
<td>Heterotrophs do not make their own food.</td>
<td>Heterotroph</td>
<td>1/17/2012 Priya: Negative information</td>
<td>1/11/2012 Shivani: Reviewed</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Example analysis showing planning and encoding at sentence level. Two universal truths were generated, one which was able to be encoded, and the other that presents a KR Issue.

<table>
<thead>
<tr>
<th>Chapter 10: PHOTOSYNTHESIS</th>
<th>Sentence</th>
<th>Non-Critical Knowledge</th>
<th>Relevant Knowledge</th>
<th>Relevant Knowledge Encoded</th>
<th>Relevant Knowledge Not Encoded</th>
<th>Reason for Not Encoding Knowledge</th>
<th>Team Comments on Knowledge Not Encoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>The electron acceptor of the light reactions, NADP+, is first cousin to NAD+, which functions as an electron carrier in cellular respiration; the two molecules differ only by the presence of an extra phosphate group in the NADP+ molecule.</td>
<td>None</td>
<td>The electron acceptor of the light reactions, NADP+, is first cousin to NAD+, which functions as an electron carrier in cellular respiration; the two molecules differ only by the presence of an extra phosphate group in the NADP+ molecule.</td>
<td>The electron acceptor of the light reactions, NADP+, is first cousin to NAD+, which functions as an electron carrier in cellular respiration; the two molecules differ only by the presence of an extra phosphate group in the NADP+ molecule.</td>
<td>NADP+, the first cousin of NAD+, the two molecules differ only by the presence of an extra phosphate group in the NADP+ molecule.</td>
<td>Knowledge cannot be encoded in AURA</td>
<td>This is critical information. But there is limitation in AURA to represent the knowledge of the link between two chemical entities.</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Sentence-Based Coverage Analysis, in which the original sentence can be broken down into critical (relevant) and non-critical (irrelevant), as well as encoded and not encoded information. Here, part of the sentence could not be encoded due to AURA Limit

This work can be further synthesized to generate a sentence-based coverage analysis, such as shown below, in Table 2:

The example analysis output for a daily report (Table 3) shows the extensive categorization of content based on UTs. The asterisk indicates that nine UTs are
pending being written, and the parenthetically added values indicate the number of sentences from the textbook sentences and the figure sentences:

<table>
<thead>
<tr>
<th>All Sentences</th>
<th>Relevant Sentences</th>
<th>Total UTs for encoding</th>
<th>Action items made</th>
<th>Action items not made</th>
<th>Encoded knowledge</th>
<th>Already present</th>
<th>KR Issues</th>
<th>KE Issues</th>
<th>JIRA issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>875</td>
<td>335 (253 + 82)</td>
<td>722 (582 + 140*)</td>
<td>437</td>
<td>285</td>
<td>82</td>
<td>163</td>
<td>46</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Categorization of Content Based on UTs.

Historically, coverage analysis is presented at the sentence level, in a table similar to Table 3. This information can also be used to generate a graphical representation, as shown in Figure 4, below. The meaning of the various groups of sentences is described in Table 4.

![Figure 4: Chapter 10 Graphic for Coverage Analysis Statistics in 2010.](image)

<table>
<thead>
<tr>
<th>Taxonomy</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Sentences</td>
<td>Sentences that were analyzed during coverage analyses</td>
</tr>
<tr>
<td>Critical Sentences</td>
<td>Sentences that are relevant to answer Biology questions</td>
</tr>
<tr>
<td>Critical Sentence – CK</td>
<td>Sentences that are computable and are relevant to answer Biology questions</td>
</tr>
<tr>
<td>Non-critical Sentences</td>
<td>Sentences that are not relevant to answer Biology questions/general statements</td>
</tr>
<tr>
<td>Non-critical Sentences – GK</td>
<td>Sentences that are general statements and are thus not relevant</td>
</tr>
</tbody>
</table>
Table 4: Sentence Groupings in Coverage Analysis.

<table>
<thead>
<tr>
<th>Non-critical Sentences - AP</th>
<th>Sentences that are not relevant to answer Biology questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redundant Sentences</td>
<td>Sentences that contain repeated information</td>
</tr>
<tr>
<td>Encoded Sentences</td>
<td>Sentences that are critical and have been encoded</td>
</tr>
<tr>
<td>Semi-Encoded Sentences</td>
<td>Sentences that are critical and have been partially encoded</td>
</tr>
<tr>
<td>Sentences not encoded</td>
<td>Sentences that are critical and have not been encoded completely or partially</td>
</tr>
<tr>
<td>Sentences not encoded due to AURA limitations</td>
<td>Sentences that are critical and have not been encoded due to AURA limitations</td>
</tr>
<tr>
<td>Sentences not encoded due to KR issue</td>
<td>Sentences that are critical and have not been encoded due to knowledge representation issues</td>
</tr>
</tbody>
</table>

**Diagram Analysis Statistics.** The diagram analysis is followed by diagram statistics analysis, which is further categorized into ‘detailed statistics analysis’ and ‘summary statistics analysis’. The detailed analysis contains the following parameters:

1. Chapter Name
2. Diagram Number
3. Total Lines (Inclusive of irrelevant sentences in caption)
4. Total Biology Relevant Lines
5. Noncovered Biology Relevant Lines
6. % of Biology Relevant Knowledge not covered in text

The current KE process includes encoding of all Biology Relevant content of the figures, using the same approach for Biology Relevant sentences. Statistics about encoding of figures are rolled into the coverage analysis statistics above, inclusive of textbook sentences and diagram-associated sentences, and their corresponding UTs.

For Chapter 6, Diagram 5 (6.5), the diagram analysis is shown in **Table 5**, below. It includes sentences directly indicated by the diagram, and sentences that convey the same information in the text, when they occur.

<table>
<thead>
<tr>
<th>Figure #</th>
<th>Page</th>
<th>relevant knowledge in the figure</th>
<th>knowledge referenced in the text</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>98</td>
<td>A prokaryotic cell lacks a true nucleus and the other membrane-enclosed organelles of the eukaryotic cell and is therefore much simpler in structure.</td>
<td>not covered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prokaryotes include bacteria and archa; the general cell structure of the two domains is essentially the same.</td>
<td>Organisms of the domains Bacteria and Archea consist of prokaryotic cells.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A typical rod-shaped bacterium has these parts: pili, nucleoid, ribosomes, plasma membrane, cell wall, capsule, flagella.</td>
<td>not covered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In prokaryotes, the pili are on the outside of the capsule.</td>
<td>not covered</td>
</tr>
</tbody>
</table>
In prokaryotes, the pili cover the surface of the capsule.

In prokaryotes, flagella can be more than one in number.

Flagella are longer in length than pili.

The plasma membrane is inside the cell wall.

In prokaryotes, the cell wall is inside the capsule.

All other parts of the cell are inside the plasma membrane.

Pili are attachment structures on the surface of some bacteria.

Nucleoid is a region where the cell’s DNA is located (not enclosed by a membrane).

Ribosome is an organelle that synthesizes protein.

Plasma membrane is a membrane enclosing the cytoplasm.

Cell wall is a rigid structure outside the plasma membrane.

Capsule is a jelly-like outer coating of many bacteria.

Flagella are a locomotion organelle of some bacteria.

Fimbriae are attachment structures on the surface of some prokaryotes.

In a prokaryotic cell, the DNA is concentrated in a region that is not membrane-enclosed, called the nucleoid (Figure 6.5).

Ribosomes, which are complexes of ribosomal RNA and protein, are the cellular components that carry out protein synthesis (Figure 6.10).

All cells have several basic features in common: They are all bounded by a selective barrier, called the plasma membrane. Inside all cells is a semifluid, jelly-like substance called cytosol, in which subcellular components are suspended.

Table 5: Diagram Analysis for Diagram 6.5. Note that relevant knowledge in the figure is referenced in the text in only some of the cases.

Diagram-related sentences are grouped together, and calculation is done to indicate percentage of relevant knowledge conveyed by a diagram that is not covered in the text. An example of these statistics is shown in Table 6.

<table>
<thead>
<tr>
<th>Figure # (Chapter . Diagram)</th>
<th>Page</th>
<th>Number of All Lines</th>
<th>Number of Biology Relevant Lines</th>
<th>Non-Covered Biology Relevant Lines</th>
<th>Percent of Biology Relevant Knowledge Not Covered in Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>98</td>
<td>18</td>
<td>18</td>
<td>14</td>
<td>77.8</td>
</tr>
</tbody>
</table>

Table 6: Diagram Statistics for Figure 6.5. 77.8% of the relevant knowledge is not covered in the text.
2.4.2 Regression Testing

This section will be written after we are done implementing the regression testing. It will explain our regression testing process, and how to interpret the scores that are reported.

2.4.3 Reporting Crashes

If the system hangs due to AURA crash or some other external software issue, then report the error along with AURA log files and screenshot of the issue. Restart AURA and resume debugging.

Using the AURA Log report for debugging: The AURA log report is generated every day and logs issues such as AURA crashes or freezes. These issues can occur while saving or modifying a concept map. You can share this log file with the SRI team members, who can help in locating any encoding problem associated with the concept map. Include the relevant part of the log, if one exists, with any bug report or JIRA issue that is filed.

3 Collaboration Processes

We expect to have at least two or three biologists involved in the KB content creation. Therefore, we define a process for them to work together as a team. This process is based on the pilot experiment conducted during 2009.

In the Pilot Phase, the KB team implemented a collaboration model where each SME played two roles: as a contributor and as an integrator.

**Contributor:** A KFE playing this role is responsible for working from the textbook and formalizing it using AURA.

**Integrator:** A KFE playing this role is responsible for merging two knowledge bases created by two contributors. The integrator judges the quality of the contributions by inspection, questions, answering, and discussion with the team.

When working as a contributor, each SME works on that SME’s own machine. Then, the contributions of all the SMEs are integrated into a common KB on a single dedicated machine. For example, a contributor SME who performs knowledge entry on that contributor’s own machine may then perform the integration by repeating the same task on the integration machine. We now describe this collaboration model in more detail.

3.1 Planning and Allocation

The first step in the process is planning and allocation. Each of the SMEs studies the knowledge to be represented (e.g., page, section, chapter, or UT concept) of the Biology textbook and determines if any section requires discussion. Key concepts
that appear frequently should be identified and discussed. The team then divides the knowledge to be encoded among the SMEs, based on the following guidelines.

- Sections or concepts that are closely related to each other are assigned to only one SME. The SME can apply this SME’s own understanding of prior sections to the following related sections, making the process more efficient.
- Several concepts are complex and are associated with various processes. These frequently used concepts are detected in the initial planning phase, and one SME is assigned to do initial planning of all such concepts. In some cases, a detailed analysis of these concepts is required, including corpus analysis, identification of exceptions, taxonomic analysis, and modeling review. For instance, pre-planning was required to determine what distinctions between Anaerobic Respiration, Aerobic Respiration, Cellular Respiration, and Fermentation were appropriate with respect to *Campbell Biology*, and how Cellular Respiration might be reified, knowing that there were variations in these processes with respect to the organism type (prokaryotic or eukaryotic), starting molecule (traditionally modeled with glucose), alternate pathways (such as cellular respiration of fatty acids), and relationships between processes (e.g. if glycolysis would be considered as a step of both or either cellular respiration and fermentation, or a precursor of it). Identification of this as an area of pre-planning need was indicated by the following: (a) the text directly indicates conflicting perceptions (e.g. *Campbell Biology* presents multiple views by scientists that are all valid, but that *Campbell Biology* adopts one), (b) models with exceptions are used in teaching about this topic (e.g. cellular respiration of glucose by a eukaryote is used as the primary basis for teaching in *Campbell Biology*), and (c) the content is very complex (for instance, with multiple steps, relationships to multiple core themes, and variation based on contexts).

- Some chapters are dependent on others for background knowledge. Chapter dependencies have been identified by SRI Biologists and can be used in the planning process. These dependencies are found at the KE Process page of the Halobook Website (http://www.ai.sri.com/halo/halobook2010/KE-process.html).

### 3.2 KB Development

In their role as contributors, the SMEs undertake knowledge entry through the following steps, as shown in Figure 5:

- Each SME encodes the relevant information for the sections assigned to that SME.
- Each SME tests the encoded knowledge in that SME’s private KB.
- Once satisfied with the knowledge developed, that SME then integrates changes into the common KB on the integrator machine. While integrating the KBs, the encoded knowledge is tested again by posing test questions.
Once the integration is complete, the contributors replace their private KBs with the integrated KB, and perform their new changes by starting from the integrated KB.

**Open Issues with our collaboration process:**

- Our process for Planning Content Creation specifies reviewing sketches of the key concept maps before the SMEs start encoding. It is an open question whether there should be a review of the concept maps as each contributor is creating them. How often should such a review be? It will vary depending on the ability and experience of the SME.
- How frequently should the integration be done?
4 Commonly Occurring Knowledge-Engineering Problems

The knowledge-engineering issues mentioned here are not exhaustive, and we will add to this discussion as new issues are encountered.

4.1 Functions

Functions are understood as special events in which an entity participates. Alternatively, it is a reason for the entity’s existence. In the following sections, we present guidelines for identifying functions in the text, and how to encode them on a Concept.

4.1.1 Encoding Functions

Functions are understood as special events in which an entity participates. Alternatively, it is a reason for the entity’s existence. The decision of whether or not an event is a function is made by an SME, based on knowledge of Biology. This is achieved by simply exposing the slot “has-function” during encoding.

![Diagram](image)

**Figure 6: An inherent function of the Golgi apparatus – “to store chemicals”**

During encoding, the SMEs distinguish between two sorts of functions. Inherent functions are those that always hold of an entity, regardless of where it occurs. Such functions are encoded on the Concept of the entity. **Figure 6** shows an inherent function of a Golgi apparatus – “to store chemicals”. Contextual functions are performed by an entity in the context of another entity or process. **Figure 7** shows the function of a smooth endoplasmic reticulum (ER) in the context of a liver cell – “to detoxify drugs and poisons”.

36
The new approach to functions has two benefits. It eliminates imprecise answers by giving the SME control in specifying which events are functions. It leads to more natural conceptualizations, because functions are not just restricted to the Concept of the entity in question, but can in principle appear anywhere in the KB. This alleviates (but does not entirely solve) the problem of incompleteness in answers to questions about function.

### 4.1.2 Identifying Functions in the Text

Function is presented in a wide range of ways. Presentation of function alongside conjecture, over-generalization, or over-specialization of context can make encoding function difficult. The following guidelines were developed to ease encoding function using the has-function slot. Below are also several examples along with solutions for how they would be encoded.

1. Encode function slots when function is overtly used in text. Figures should be considered as well.

   “Membrane carbohydrates function as markers that distinguish one cell from another.”

   Solution: Encoding should be on the Membrane-Carbohydrate Concept. Mark could be interpreted here as a role or a function, and Distinguish could be interpreted as a function. Based on their definitions and subclasses, Distinguish is a more accurate function (subclasses include identification, comparing, and recognition).
2. Encode function slots when function is used generally, and inspection of text surrounding this use indicates the process to which that function should be mapped. Sometimes this process will be described rather than named, in which case creating a name (class) for the function is appropriate (and its description would be mapped here).

“Channel proteins function by regulating movement of molecules across a membrane.”

Solution: Encoding should be on the Channel-Protein Concept. Function is Regulation, which exists in the KB and has the subclass Control, which is implied here. The definition for Regulation fits within the larger picture that the Channel Protein regulates balances of molecules on either side of a membrane; its definition is “The act of bringing uniformity; making regular is regulation.”

3. Role is often used as an analogue of function. Variations of the role term can be used to encode the function.

“Roles of Cytoskeleton: Support, Motility, and Regulation”

Solution: Encoding should be done on the Cytoskeleton Concept. Functions would be Support, Move (Motility does not exist, and Move’s definition matches the use here and its subclasses include some types of motility that are related to this text, such as Locomotion), and Regulation. This may be a case where an over-generalization may be made, which could be problematic. If not every cytoskeleton has these three functions, then the has-function relation could be used in specific subclasses of cytoskeleton or in contextual events or entities for which that specific function is realized. See “Where to Encode Function” below, for more details.

4. Encode function when specific entities/regions have a role or activity that is true for all entities/regions of this type. Functions often precede extensive presentation of them with examples of varying cases.

“Hydrophobic core impedes direct passage of ions.”

Solution: Encoding should be in Hydrophobic-Core Concept. Impede does not exist in the KB, but Block does. Block’s definition matches the use of Impede here, so it should be used as the encoded function.

5. If a function is presented and no exceptions are given, the function can assume to be generalized for all instances.

6. Some information is presented hypothetically for historical purposes, not because it is conditional. Many of these ideas have not been confirmed with substantial evidence, but these hypotheses agree with general understanding of functions and structures. These functions should be encoded.
“Possible cytoskeletal function is regulation of biochemical activities.”

Solution: Cytoskeleton is the entity to be encoded. “regulation of biochemical activities” is not specific, so “regulation” is sufficient to describe this sentence. If this topic comes up later in the text with more detail, this event could be specialized.

7. If a function is given, and a specific example provided additionally, then the function can be generalized, and the specific details can also be encoded in a specific instance of this general function.

“Smooth ER function includes detoxification of drugs and poisons.” General function for Smooth ER
“Barbiturates are examples of drugs detoxified by the smooth ER in liver cells.” Specific function for Liver Cell Smooth ER

Solution: The first sentence would be encoded on the Smooth Endoplasmic Reticulum Concept. The function would be Detoxification of the objects Drugs and Poisons. The second sentence would be encoded on the Liver Cell Concept. Liver Cell has part Smooth Endoplasmic Reticulum. This specific Smooth Endoplasmic Reticulum would inherit the function of detoxifying drugs from the Smooth Endoplasmic Reticulum class. This Liver Cell Smooth Endoplasmic Reticulum could have the function of Remove objects Barbiturates, which should be unified with Drugs (a barbiturate is a specific drug) from the inherited Detoxification. A further relationship could be made between the Detoxification and Remove events.

8. When structure is indicated as “playing a crucial role in” or “is important in” an event, this may imply a function of that structure.

“The Role of Membrane Carbohydrates in Cell-Cell Recognition (title). Cell-cell recognition, a cell’s ability to distinguish one type of neighboring cell from another, is crucial to the functioning of an organism. It is important, for example, in the sorting of cells into tissues and organs in an animal embryo.”

Solution: Three clues are here that has-function should be encoded for Membrane-Carbohydrate. The function here is clearly Cell-Cell Recognition. In evaluating the existing function of Membrane-Carbohydrate according to example (1) above, the general function Distinguish could be specialized if the SME felt this specialization was redundant with the Distinguish function.

9. If the function is given only within a specific context, then the function should be assigned to only that specific instance.

10. Some entities will have no function that is always true of that entity (on a textbook relevant level) level.)
11. The review at the end of each chapter often provides a summary of functions. The supplementary materials from Pearson such as the Study Guide also provide a place to look for functions. Some are over-generalized to all instances of an entity, so screening these lists should be used as a resource to check for prior encoding of function.

**Anthropomorphized Function**

The knowledge base includes content that is true for biology in nature, at the scope of advanced high school/early college biology. So what do we do with anthropomorphized or human-directed application of biology including addition of new function, perception, and unique laboratory contexts?

Anthropomorphized function is treated the same way as other functions; has-function is used in both cases. Similarly, there is no taxonomic distinction in the knowledge base between a “natural” role and a human-imposed role. In both of these cases, context can be established with respect to the event that is the function or in which a role is realized. For instance, scientists are developing new biofuels and carburetion fuels derived from biological materials. Plant crops can be harvested and used in manufacture of ethanol, for instance, so plants, in this context, are a source of biofuel. Because this role (Biofuel) is only relevant to a plant within the context of manufacturing to generate biofuel, the role is intentionally not encoded on the Plant Concept, and instead is encoded on Concepts such as Bioethanol-Production, which is a subclass of Biofuel-Production.

**Use Case for Establishing Correct Terminology for Function**

Suppose we are encoding “water functions to split molecules, forming acids and bases.” This simplifies to the phrase “water splits”. The word “split” can have various connotations, such as “break”, “take-apart”, and “divide”, or in this case, we might be interested in encoding this as “hydrolysis.”

Hydrolysis does not occur in our eleven-item functional vocabulary, as it is far too specific. Hydrolysis can be generalized to be a type of “separate”, however, which is one of our eleven functional terms. I should confirm this by looking at AURA before I encode this.

Hydrolysis does not occur immediately below “separate” in the taxonomy, so I can look through progressively more specific events beginning from “separate”, or I can search for “hydrolysis” in the KB.

Looking at the details pane for “hydrolysis” there is indication in its definition that a second molecule breaks as well. I see in the text I am reading that water is splitting another molecule, so this makes sense. Hydrolysis is a type of break, and break is a type of separate, so it is appropriate to use this term as a function.

A function can also be presented in the context of an event. For example, “Enzymes
catalyze oxidation during cellular respiration” would be encoded on the Concept of Cellular-Respiration.

4.2 Facilitating Functions

Certain superstructures have substructures (inclusive of parts or regions) that are specifically utilized to accomplish the function of the superstructure. For example, Chlorophyll-A contains the functional group CH3. The text describes how CH3 facilitates Chlorophyll-A's function of absorbing violet-blue light. We want to show this relationship in our concept maps. A general mapping showing how facilitates is used between a substructure and a function is shown in Figure 8. Figure 9 shows the specific mapping for Chlorophyll-A's porphyrin substructure.

![Diagram showing facilitation relationship](image)

**Figure 8:** A general mapping showing how facilitates is used between a substructure and the function of a superstructure.
Create a facilitates relationship whenever the text explicitly describes a substructure aiding or participating in the function of a superstructure. A facilitates relationship is always between an Entity, such as porphyrin, and an Event, such as Absorbing Violet-Blue Light. A facilitates relationship always originates at the Entity and terminates at the Event. A facilitates relationship can occur only in the context of a function, so the chosen Event should already be related to some other concept with a has-function relationship.

Note that a function may have many facilitates relationships associated with it. For example, consider Figure 10. The text describes how the hydrophilic channel and the binding site both enable Membrane’s function. Since both structures are important to the function, we create two relations: (1) Hydrophilic Channel facilitates Move-Into and (2) Binding Site facilitates Move-Into.

Figure 9: Concept showing how the porphyrin ring facilitates the function of chlorophyll
Note that you may create facilitates relations at multiple levels of abstraction for a given function. There is no limitation to what structures may facilitate a function. If the text describes them as being especially important, a facilitate relationship is warranted. For example, consider Figure 11. After reading a bit more, we came across a passage where the text described Porphyrin being important to the function of Chlorophyll-A. Even though it is a different level of abstraction than CH3, we want to encode all the important information described by the book, so we create the relationship Porphyrin facilitates Absorption of Violet-Blue Light to encode this fact.
Where to Encode a Facilitates Relation

Encode the facilitates relation in the concept map where the relevant has-function relationship occurs. For example, let us say that Chlorophyll-A has-function Absorption of Violet-Blue Light, and that relationship is encoded in the concept map for Chlorophyll-A. You want to encode that CH3 facilitates this function. Following our rule, you should encode CH3 facilitates Absorption of Violet-Blue Light in the concept map for Chlorophyll-A.

It is important to remember that facilitates occurs only in the context of a function. If you find yourself in a situation where a facilitates relationship seems warranted but the event does not already have a has-function relationship, this is an opportunity to consider if the event is the function of some structure. You are most likely making a mistake if you encode a facilitates relationship to an event that is not a function.

You may want to encode a facilitates relationship between two events. Remember that other relationships exist to describe how one event supports another event. For example, consider causes, inhibits, leads to, enables, and subevent.

Use Case

You have seen how the facilitates relation can exist in the context of Chlorophyll-A and Membrane. For a concrete example, say that we want to encode the sentence,
“In eukaryotic cells, mitochondria house most of the metabolic equipment for cellular respiration.”

Earlier in our knowledge engineering process, we encoded that Eukaryotic Cell has-function Cellular Respiration, a process that harvests energy. Now we want to describe how mitochondria facilitate this function. To do this, we create a new relationship: Mitochondria facilitate Cellular Respiration.

Notice the key phrase that led us to use the facilitates relation: “mitochondria house most of the metabolic equipment for cellular respiration.” This phrase tells us that mitochondria are especially necessary for cellular respiration to occur because they have equipment that the process needs. We might later learn specific substructures of mitochondria that facilitate the function; this is correct. Remember that you are allowed and encouraged to create multiple facilitates relations at many levels of abstraction. Whenever a structure is especially involved in the accomplishment of a function, it facilitates that function.

In fact, that is exactly what happens. We continue reading about Cellular Respiration and come across the following sentences. “The Krebs cycle and the electron transport chains are located inside the mitochondria.” And “The electron transport chain converts the chemical energy to a form that can be used to drive oxidative phosphorylation, which accounts for most of the ATP generated by cellular respiration.”

Aha! The text has identified a substructure of mitochondria, the Electron Transport Chain (later described as “a collection of molecules embedded in the inner membrane of the mitochondrion”), that plays a significant and important role in Cellular Respiration. Let us encode that relationship: Electron Transport Chain facilitates Cellular Respiration.

Notice how we identified this relationship. The text says that the electron transport chain “accounts for most of the ATP generated by cellular respiration.” A structure that significantly contributes to the outcome or accomplishment of a function is most likely a structure that facilitates that function.

As we learn more about how this structure interacts with Cellular Respiration, we may want to identify a specific function for the Electron Transport Chain and relate that subfunction to the superfunction of Cellular Respiration. This is appropriate. Even though we have more information about how Electron Transport Chain works within Cellular Respiration, it is still true that Electron Transport Chain facilitates the function.

4.3 Qualitative Relations

Guidelines for Identifying Qualitative Relations Presented in Text
Qualitative relations describe how one property affects another property. We call them “qualitative” because they are generalized causal descriptions about positive/negative change rather than specific, quantified formulas or use of specific quantified values. For example, we might say that the duration of rainfall positively influences the amount of water of a lake. We can talk about the amount of water being low, normal, high, and so on, but no quantification (e.g., 10 feet) is necessarily known. The qualitative relation “positively influences” implies that while rainfall occurs, the amount of water in the lake increases.

As an aside, what if we had the required knowledge to define quantitative relations in the rainfall example? Let us say we knew that while rain occurs, the water level rises by 10 liters every minute. We could write a formula \( f(t) = t \times 10 \) liters. Recall that formulas such as this are encoded using equations in Concepts. Relationships between quantitative values can also be encoded using tables in Concepts.

Use a qualitative relation whenever the text talks about the non-numeric value of one property or the rate of change in one property effecting the non-numeric value or rate of change in another property. A qualitative relation can also be used to abstract from examples, where cardinal values might be used, when the qualitative relationship between properties is relevant to the textbook content, and the specific example is trivial (in other words, the example is provided in order to teach [a generalization] by example).

Note that, although many qualitative relations are between properties of different concepts (such as in the rainfall and lake example), this need not be the case. For example, we could create the following qualitative relation: the height of a person is directly proportional to the size of the person.
Figure 12: Concept map for a membrane that shows qualitative relations.

Also note that a single property may affect or be affected by multiple other properties. For example, consider the following concept map about a membrane in Figure 12.

The text describes how the rate of phospholipids moving apart and the rate of phospholipids moving together (compacting) both affect the fluidity of a membrane. We encode both relationships: (1) Rate of Move-Apart is directly proportional to the Fluidity of Membrane, and (2) Rate of Move-Together is inversely proportional to the Fluidity of Membrane. Although it looks complex, both relationships are necessary to capture the complex interaction that these two events have on the fluidity of the membrane.

Remember, you need not restrict yourself to a single qualitative relation if the text describes a complex interaction between properties.

Choosing a Qualitative Relation
There are five qualitative relations from which to choose. Remember that qualitative relations always relate two properties to each other. Also, qualitative relations are always unidirectional, even Correspondence. If you want to make them bidirectional, remember to create two relations (one for each direction).

**Influencing Relationships**

- **Positive influence**
  - When the source’s property value is positive, the target’s property value is increasing.
  - When the source’s property value is negative, the target’s property value is decreasing.
  - When the source’s property value is zero, there is no implied change in the target’s property value.
  - Example: Quantity of inflowing water in a container positively influences the level of water in that container

- **Negative influence**
  - The inverse of positive influence.
  - When the source’s property value is positive, the target’s property value is decreasing.
  - When the source’s property value is negative, the target’s property value is increasing.
  - When the source’s property value is zero, there is no implied change in the target’s property value.
  - Example: Quantity of out-flowing water in a container negatively influences the level of water in that container.

**Proportionality Relationships**

- **Directly proportional**
  - When the source’s property value is increasing, the target’s property value is increasing.
  - When the source’s property value is decreasing, the target’s property value is decreasing.
  - When the source’s property value is zero, there is no implied change in the target’s property value.
  - Example: Downward force on a free-falling object is directly proportional to its mass

- **Inversely proportional**
  - The inverse of directly proportional.
  - When the source’s property value is increasing, the target’s property value is decreasing.
  - When the source’s property value is decreasing, the target’s property value is increasing.
  - When the source’s property value is zero, there is no implied change in the target’s property value.
  - Example: Gravitational force between two objects is inversely proportional to the distance between the two.
Correspondence

- Correspondence
  - The source’s property value implies the same value in the target’s property.
  - Or if explicitly specified, a certain value in the source’s property implies another certain value in the target’s property; they need not be the same value.
  - Correspondence differs from the other four relations because it does not describe the kind of causal relationship between the two properties. This relationship simply describes that two values correspond and leaves absent how that occurs.
  - Example 1: When the rainfall is high, the level of river water is high.
  - Example 2: When the switch is on, the amount of electricity is high.

Sometimes, it is hard to choose between an influencing relationship, a proportionality relationship, and a correspondence relationship. Here are some guidelines and examples to help you decide:

- If a change in one thing causes a change in another, use a proportionality relationship.
  - “For example, the average body size of many North American species of birds and mammals increases gradually with increasing latitude.”
    - This can be simplified to “The average body size increases gradually with increasing latitude.”
    - Recall the rule above for Direct proportionality: “When the source’s property value is increasing, the target's property value is increasing.”
    - The average body size of North American birds is directly proportional to the latitude.
  - “Average size of yarrow plants (*Achillea*) growing on the slopes of the Sierra Nevada mountains gradually decreases with increasing elevation.”
    - The size of yarrow plants is inversely proportional to the elevation.

- If the certain value of one thing causes a change in another, use an influencing relationship.
  - “The acidification of the wall activates enzymes called expansins that break the cross-links (hydrogen bonds) between cellulose microfibrils, loosening the fabric of the wall. (Expansins can even weaken the integrity of filter paper made of pure cellulose.)”
    - The activation of enzymes has a positive influence on the loosening of the wall fabric.
  - “When light triggers the rhodopsin signal-transduction pathway by altering retinal, an enzyme converts cGMP to GMP, which disengages from the Na+ channels (see Figure 49.13 in the Campbell Biology
textbook) This closes the channels, decreasing the membrane’s permeability to Na+ and hyperpolarizing the membrane potential.”

- The state of the channels (normal or closed) has a negative influence on the permeability of the membrane, such that when the channels are in state = closed, the permeability of the membrane decreases.

- If you know that the value of one property affects the value of another property but are not sure how, use a correspondence relationship.
  - “When a membrane solidifies, its permeability changes…”
    - The solidity of a membrane corresponds to the permeability of the membrane.

**Use Case**

Suppose we are encoding the sentence, “When a membrane solidifies, its permeability changes, and enzymatic proteins in the membrane may become inactive.” This sentence breaks down into two potential qualitative relations: “When a membrane solidifies, its permeability changes” and “When a membrane solidifies... enzymatic proteins in the membrane may become inactive.” Let us look at each in turn.

Unfortunately, the phrase “when a membrane solidifies, its permeability changes” does not tell us the nature of the qualitative relation. We could rephrase this as “the solidity of a membrane has an unknown causal relationship with the permeability of that membrane.” Because we do not know the nature of the relationship, we should draw a Correspondence relationship from the solidity property of membrane to the permeability property of membrane. Recall that this relationship implies that a change in the first property (solidity) causes a change in the second property (permeability), but does not describe how that change comes about.

“When a membrane solidifies... enzymatic proteins in the membrane may become inactive” is clearer. We may rephrase the sentence to the following: “as the solidity of a membrane increases, the activity of enzymatic proteins contained in that membrane decreases.” Another way to say this is that “the solidity of a membrane is inversely proportional to the activity of enzymatic proteins contained in that membrane.” As a KE, you would draw an Inversely Proportional relationship from the solidity property of the membrane to the activity property of enzymatic proteins.

**Where to Encode a Qualitative Relation**

Once a qualitative relation has been identified, the concept map in which it will be encoded must be determined.

You should encode the qualitative relation in the concept map where it is universally true, which is usually given in the sentence itself (as with the example
below) or can be inferred by considering the overall chapter or section topic as the context for the information. Because property change usually occurs in the context of an event and qualitative relations describe changing property values, it is often correct to decide that a qualitative relation exists in the context of some event or process (rather than being universally true of an entity) and to encode the relation wherever that event or process was encoded.

For example, consider the following sentence: “Average size of yarrow plants (Achillea) growing on the slopes of the Sierra Nevada mountains gradually decreases with increasing elevation.” This sentence implies a relationship between the size of yarrow plants and the elevation of yarrow plants. However, the sentence allows us to infer only that the relationship holds for yarrow plants on the Sierra Nevada mountains. As a result, we should encode Elevation of Yarrow Plants is inversely proportional to the Size of Yarrow Plants in the concept map for the Yarrow Plant At Sierra Nevada Mountains. If we instead encoded this relationship in the concept map for Yarrow Plants, our encoding would imply that the qualitative relation holds for all yarrow plants regardless of location. This is not true of all Yarrow Plants, so should not be encoded on the general Yarrow Plant Concept.

4.4 Roles
This document provides guidelines for two issues related to roles:

1. Role Identification – When should a biology term be classified as a role instead of an entity?
2. Role Use – Once a role has been identified, how should it be used?

A detailed discussion of these guidelines is available in a separate set of slides. Here, we summarize the main decisions.

4.4.1 Role Identification
We define a biological role as a temporary job or duty of an entity, including

a. Ecological roles, such as, Primary, Secondary, and Tertiary Consumers
b. Participants in an electron transport chain, such as Electron Carriers, Electron Donors, and Electron Acceptors
c. Nutrients, such as Limiting Nutrients, Micro/Macro-nutrients

Based on this definition, it is sometimes confusing to determine whether a term should be classified as a role or an entity. To decide, use the following procedure:

I. A concept is an entity if
   a. It has a structural description
   b. It can be measured or drawn
   c. Examples – Cell, Biomembrane
II. A concept is a role if
III. Roles for life – Some concepts are defined in terms of participation in an event, but persist for an entity’s lifetime:
   a. By default, we treat such cases as Entities – Autotroph, Heterotroph
   b. However, if it is closely related to a concept that satisfies the definition of a role (by clause II), then it should be treated as a role
      i. Parent is a role that can be acquired
      ii. Daughter is a role for life
      iii. It does not make sense to treat parent and daughter differently, and so Daughter should be classified as a role

4.4.2 Role Definition and Use

Currently, there are two methods of using roles in the KB:
   a. Unification – The role is unified with the entity that plays it
   b. Plays – The slot plays and in-event are used to relate the role to the entity and the event

We will be switching to the “plays” approach. For example, the role Electron-Carrier should be defined as shown in Figure 13.

![Figure 13: Concept of Electron-Carrier-Role. The relevant event should be included in the Concept.](image)

To say that nicotinamide adenine dinucleotide phosphate (NADPH) and Water play the role of electron carriers in photosynthesis, we use these roles on the Concept of Photosynthesis, as shown in Figure 14. In addition, if we want to indicate that the role is a function, then the has-function slot should be used in addition to the role. (Use of these two Carry events as subevents of photosynthesis may not be ideal for modeling the collection of subevents in Photosynthesis, but the presentation of roles here is useful for exemplifying how roles may be encoded.)
4.5 Chemistry Representation: Elements and Atoms, Reactions, Catalysts, Biomolecules, and Bonds

**Elements and Atoms**
Carbon element is a collective material of carbon, in which carbon could be in several variants. These variants may include isotopes (such as C-12, C-13, and C-14) or ions, if they are relevant. Due to the ambiguous nature of the term element, we have preferred to encode carbon in a more precise way, using atoms, ions, and isotopes instead of element. In the knowledge base, Carbon exists as an atom with a defined number of protons and electrons, and subsequently an atomic number. Its atomic weight is not specified, nor is its number of neutrons, as this is based on the number of neutrons it has. Carbon-Isotope is a subclass of Carbon (atom) that has neutrons. Carbon-12, Carbon-13, and Carbon-14 exist as subclasses of Carbon-Isotope in which the number of neutrons and subsequent atomic weight is defined.

**Chemical Reactions**
Figure 15 shows the new concept Chemical-Reaction. The raw-material and result are now constrained to be Chemical-Entity, which is a count term, but can also include energy as shown in Figure 16. As a byproduct of this definition, we have guided the SMEs to use count terms (molecules and not substances) uniformly in the encoding of reactions.

![Figure 15: New encoding of Chemical-Reaction. The raw-materials and results are now Chemical-Entity, which are count terms.](image1)

![Figure 16: New encoding of exergonic reactions, which result in energy.](image2)

Evaluation of encoding of chemical reactions revealed use of domain-general subevents such as Create, in lieu of domain-specific subevents such as Synthesis-Reaction. We have guided SMEs to use domain-specific terms when possible, just as encoding should be on the most specific level that is relevant, when possible.

Related to chemical reactions are metabolic pathways and metabolism events, described below.

**Metabolism** is defined as the totality of an organism’s chemical reactions, grouped into at least one metabolic pathway with additional subevents. Because metabolism
includes at least one metabolic pathway, it is not classified as a chemical reaction, but instead as a Cellular-Process. Metabolism is subdivided into anabolism and catabolism. Anabolism (building or synthesis metabolism) includes photosynthesis. Catabolism (breaking metabolism) includes cellular respiration, anaerobic respiration, and fermentation. Note that when classifying metabolism types, the overall perception of a process as building or breaking was considered. For instance, cellular respiration includes in its subevents some synthesis reactions, but in total, it is considered a breaking reaction, as glucose breaking is traditionally tracked in this process.

**Metabolic Pathway** is defined as a series of chemical reactions that happen in a unit of time or space. In contrast, the metabolism process cellular respiration includes four metabolic pathways: glycolysis, pyruvate oxidation, citric acid cycle, and oxidative phosphorylation, that may be separate by space (glycolysis occurs in the cytosol, pyruvate oxidation may occur in the mitochondria) or time (there is a regulatory step to determine, based on oxygen gas concentration, if cellular respiration will continue after glycolysis, or if fermentation will occur, instead). All metabolic pathways require at least two chemical reactions or steps in which a chemical or chemical entity are modified, so are classified as Multi-Step-Reaction, which is a subclass of Chemical-Reaction. Metabolic pathways are also classified as either anabolic pathways or catabolic pathways. Anabolic pathways include the Calvin Cycle and Synthesis of Protein. Catabolic pathways include Citric Acid Cycle and glycolysis.

**Catalysts**
Reactions are often aided by catalysts, which include temperature, salts, and enzymes. While in science, enzymes are often involved in a reaction, they are not required, so it is not appropriate to think of an enzyme or any catalyst as an agent in a chemical reaction. Instead, the Enzyme plays the Catalyst role in the chemical reaction, and the enzyme need not have a direct relationship to the chemical reaction. **Figure 17** below shows encoding for Enzymatic-Reaction, which has a trigger: Catalyst in-event Enzymatic-Reaction and Enzyme plays Catalyst.

![Figure 17: Concept map for Enzymatic Reaction. Note that the trigger utilizes both relations and non-root nodes.](image-url)
**Biomolecules**

Biomolecules are defined as biological molecules, which are of some utility to living organisms, or could be produced by living organisms. Large biomolecules were once used prevalently as a group associated with the four main kinds of organic molecules of value to all life: proteins (with amino acid monomers), carbohydrates, lipids, and nucleic acids (with nucleotide monomers). The term macromolecule is often associated with large biomolecules, but does not necessarily include all smaller lipids, carbohydrates, and monomers. Given the relative ambiguity of the qualifier “large” in large biomolecules, and the prevalence of exceptions in these four biomolecule groups that are neither macromolecules nor large, as well as recognition of more and more polymers as having significance in science and life, both terms macromolecule and biomolecule have fallen out of favor in *Campbell Biology* and College Board (authors of the AP Biology Exam). We have intentionally avoided inclusion of these terms in the knowledge base for these reasons.

**Bonds**

Chemical bonds are inclusive of weak interactions such as Van der Waals Forces and strong bonds such as Covalent Bonds. Bonds occur between chemical entities as diagrammed in Figure 18. In covalent bonds, both chemical entities are specialized to atoms. In ionic bonds, both chemical entities are specialized to ions – anion for one, and cation for the other. For hydrogen bonds, both chemical entities are specialized to atoms, and there is clear distinction to indicate that each atom is a part of a separate molecule, as hydrogen bonds only form between multiple molecules, not within molecules.
4.6 Meronymic Relations between Entities

4.6.1 Choosing a Relation

The starting point is that we are given two entities X and Y, and it is biologically correct to say “X has Y”. For example, a membrane has a hydrophobic core, and a protein has amino acids. We then apply the following procedure:

a. X has-region Y if Y is a type of Place. Places or regions are taken to be spatial entities that are not tangible. In other words, they occupy a region of space, and can be associated with properties such as area, volume, shape. But, they cannot have properties such as mass, density, color. A hydrophobic core is a region since it cannot have a mass, density, or color. A protein is not a region since it can have all these properties.

b. X material Y if Y is tangible (typically, a substance or mass term) and pervasive in X. For example, a shirt can have cotton the substance as material. But, molecules of cotton cannot form a material.

c. X has-part Y otherwise.

We have decided to phase out the relation has-basic-structural-unit, as there was no principled way to distinguish it from a part.
Biological Regions: Functional Groups, Carbon Skeletons, and Protein Structure

We have found it helpful to consider certain biological organizations as regions. Recall from section 4.6.1 that a region is defined as a Place that is a spatial entity that is not tangible. Regions can be associated with properties such as area, volume, and shape, but not properties such as mass, density, and color. We found that functional groups, carbon skeletons, and protein structures, while at first glance appeared to have parts (so not region) were better conceived as regions with parts located in the same space of the region. This was because the region itself was not defined by parts and could gain and lose parts while not losing its identity. Use of Region as a superclass to each of these also facilitated us describing overlapping regions that were helpful is describing how an entity such as an atom could be associated with multiple regions concurrently (like two functional groups), or that a protein could exist in multiple structures concurrently (like secondary and tertiary structure).

Functional groups are assemblies of atoms that are associated with molecules. When a functional group such as a phosphate group is removed from a molecule (e.g. ATP), assembly of atoms assumes a new name – in this case, a phosphate ion. Because the functional group is only defined when attached to a molecule, it is appropriate to consider it as a region. Its constituent parts (atoms) are parts of the molecule and are located in the functional group region. Similarly, atoms occur at a carbon skeleton region, which is always a region of an organic molecule.

Protein Structure: Proteins can occur in primary, secondary, tertiary, and quaternary structures. These structures were defined as people wanted to describe a protein at varying levels of detail – first in terms of an amino acid sequence (primary structure), then a rippled or helical structure due to hydrogen bonding (secondary structure), and then due to a variety of interactions based on the hydrophilic and hydrophobic nature of each amino acid’s R group (tertiary). The quaternary structure occurs for some proteins that assume a new shape based on interaction with another molecule such as a second protein. Because all proteins have an amino acid sequence, and this sequence spatially overlaps entirely with the parts of the protein (its amino acids), it is appropriate to consider this primary structure as a region. The secondary structure overlaps entirely with the primary structure, with additional spatial and tangible features associated with this region, such as the hydrogen bonds between specific atoms that also occur in the primary structure.

Other biological regions for which has-function is appropriate include 5’ Carbon, 3’ End, and Sugar Phosphate Backbone.

4.6.2 Redundancy

We distinguish between two types:
a. Redundancy due to inherited information – This typically happens due to classes that have subclasses of very different sizes. Consider, for example, molecules, which have atoms as their direct parts. This works well for small molecules, such as carbon-di-oxide, which have atoms as direct parts. However, it leads to redundancy with macromolecules. For example, proteins have atoms as direct parts, and indirect parts, since atoms are part of polypeptides, too. Such redundancy is allowed.

b. Redundancy due to locally asserted information – An example is a protein having amino acids as a direct part, and also as an indirect part, since they are part of polypeptides, too. If there is an amino acid that is part of a protein and not part of a polypeptide, this encoding is not redundant and allowed. However, if every amino acid that is part of a protein is also a part of a polypeptide, then it is not allowed. In other words, has-part should correspond to direct parthood.

The motivation for these guidelines comes from question answering. To answer “what are the parts of X”, we wish to see only the locally asserted direct parts. Redundancy due to inherited information is harmless, whereas redundancy due to locally asserted information can cause confusion. For example, if amino acids are listed as direct parts of proteins, and polypeptides, should we interpret it to mean that there are amino acids that are part of a protein, and are not part of a polypeptide?

4.7 Superclasses vs. Subevents and Parts

During the encoding process, SMEs came across sentences where it was unclear whether to use a superclass or parthood relation (subevent or has-part). Consider the following sentences:

- Most often these existing proteins are modified by phosphorylation, the addition of a phosphate group onto the protein.
  - Is Add a superclass or subevent of Phosphorylation?
- Metabolic pathways that release stored energy by breaking down complex molecules are called catabolic pathways.
  - Is Break a superclass or subevent of Catabolic-Pathway
- Examine again the summary equation for cellular respiration, but this time think of it as a redox process.
  - Is Redox-Reaction a superclass or subevent of Cellular-Respiration?
- Recall that the plasma membrane is a phospholipid bilayer with associated membrane proteins.
  - Is Phospholipid-Bilayer a superclass or part of a Plasma-Membrane?

We now describe our guidelines for making this decision. We are given two concepts X and Y, and we need to decide whether “X is a Y” or “X has-part/subevent Y”. The following steps are used to choose:

1. Use X has-part Y if it is appropriate to say “Y is a part of X” or “the Y of X”, and
2. There is an entity Z that is a part of X and is not a part of Y, or
Note that Z may be a part in only one of the subclasses

b. More than one Y is a part of X

2. Use X subevent Y if it is appropriate to say “During X, Y happens”, and

a. There is an event Z that a subevent of X and is not a subevent of Y
   – Note that Z may be a subevent in only one of the subclasses
b. More than one Y is a part of X

We discuss some examples to illustrate the guidelines:

- It is appropriate to say “the phospholipid bilayer of a plasma membrane” and the plasma membrane has proteins as parts. These proteins are embedded in the bilayer but are not part of it. So, a Phospholipid-Bilayer is a part of a Plasma-Membrane.
- It is appropriate to say that “during phosphorylation, addition happens”. However, the textbook does not describe any other subevent of phosphorylation. So, Add is not a subevent of Phosphorylation.
- It is appropriate to say that “during active transport, molecules move through the membrane”. There are no other direct subevents of active transport. However, during active transport that uses ATP, the transport protein is phosphorylated before the movement happens. So, Move-Through is a subevent of Active-Transport.

When we conclude that Add is not a subevent of Phosphorylation, this does not mean that it is a superclass. One has to consider other possible superclasses and choose. In the case of Phosphorylation, the other candidate superclass is Chemical-Reaction. Having multiple superclasses is a way to convey different views on a concept. We distinguish between two types of views:

1. Compatible Views – Two views are compatible if the relations and properties used in one superclass are compatible with the other. For example, a Protein is a Macromolecule and an Amphipathic-Entity. A macromolecule is a big molecule. And, an amphipathic entity is one that has two hydrophilic ends and a hydrophobic core. These two views make sense with each other. For example, it makes sense to say that a macromolecule has hydrophilic ends and a hydrophobic core. And, it makes sense to talk about the size or repeating units in an amphipathic entity.

2. Incompatible Views – Suppose Phosphorylation has Add and Chemical-Reaction as superclasses. Add has the associated slots “base” and “object”, whereas Chemical-Reaction has “raw-material” and “result”. The slots “base” and “object” do not have an interpretation with Chemical-Reaction, and the slots “raw-material” and “result” do not have an interpretation with Add. This makes the two views incompatible, and one should be chosen. Given a choice between a technical and non-technical superclass, we prefer the technical one, and hence Chemical-Reaction is chosen. The relationship between Phosphorylation and Add is a KR-Issue.
4.8 Spatial Relations

We describe guidelines for using the relations is-inside, is-outside, abuts, is-at, and is-between. An encoder who wishes to use another spatial relation should (a) check that these relations do not suffice and (b) discuss the case with SRI. We begin by developing a graphical notation for the spatial extent of an entity, and then use this notation to describe the guidelines.

4.8.1 Graphical Notation for Spatial Extents

Every spatial entity has a spatial extent, which is a closed set with no holes. Examples include squares, circles, and polygons in two dimensions, or cubes and spheres in three dimensions. If an entity has holes (in the form of either pores on the surface or space in its interior), the spatial extent includes all the holes. Figure 19 shows possible spatial extents for a protein, a channel, and a cell and its membrane. Note that a cell and its membrane have identical spatial extents, because we require that a spatial extent include holes on the surface and in the interior.

![Figure 19: Examples of spatial extents for a protein, a channel, and a cell and its membrane.](image)

4.8.2 Graphical Definition of Spatial Relations

We define the relations is-inside, is-outside, abuts, and is-between graphically.

Figure 20 shows the definition of is-inside. We say that x is-inside y if the spatial extent of x is contained in the spatial extent of y. As we discussed in Section 4.8.1, spatial extents do not have holes, and so the spatial extent of x is included in that of y. Examples include “a plasma membrane is inside the extra-cellular fluid”, and “the cytoplasm is inside the plasma-membrane”. Note that the spatial extent of an entity includes the spatial extents of all its parts. So, “the phospholipid bilayer is inside the biomembrane”, and “a membrane protein is inside a biomembrane”. However, a membrane protein is not necessarily inside the phospholipid bilayer.
Figure 20: x is-inside y implies that the spatial extent of x is contained in (inside) the spatial extent of y.

Figure 21 shows the definition of is-outside. We say that x is-outside y if the spatial extents of x and y are disconnected/separate/different, but they may be touching externally. Note that “x is-inside y” does not imply that “y is-outside x”. It is inappropriate to encode that “the extracellular fluid is outside the plasma membrane”. The relation is-outside should be read as “is external to”, and it is symmetric. That is, if “x is-outside y”, then “y is-outside x”. The examples in Figure 21 illustrate the use of is-outside along with abuts.

Figure 21: x is-outside y implies that the spatial extents of x and y are disconnected, but they may be touching (externally connected).

Figure 22 shows the definition of abuts. We say that x abuts y if x and y are touching either internally or externally. In other words, there are two senses of abuts: internal abutting and external abutting. Consider the problem of spatially relating a peripheral protein to a plasma membrane. As discussed above, since the protein is part of the membrane it is inappropriate to say that “the protein is outside the membrane”. The spatial extent of an entity includes the spatial extent of all its parts, and so “the protein is inside the membrane”. It is more precise to say that “the protein is outside the phospholipid bilayer”. To indicate that it is touching, we may add that “the protein abuts the phospholipid bilayer”. If we wish to encode that the protein is on the cytoplasmic side, we add that “the protein is inside the cytoplasm”. If we wish to encode that the protein is on the extracellular side, we add that “the protein is outside the extracellular fluid”. Note that in all cases, “the protein is inside the extracellular fluid”. 
Figure 22: x abuts y if x and y are touching (either internally or externally).

Figure 23 shows a (partial) graphical definition of is-between. We say that x is-between y and z if

- x is connected to y and z. The term connected means touching or containment.
  - y is-outside z, or z is-inside both x and y

Suppose we have encoded that “the cytoplasm is inside the plasma membrane” and “the plasma membrane is inside the extracellular fluid”. This implies that “the cytoplasm is inside the extracellular fluid”. We conclude that “the plasma membrane is between the cytoplasm and the extracellular fluid”.

Figure 24: x is-between y and z if x is connected to y and z, and y is-outside/is-inside z or z is-inside/is-outside y.

4.8.3 Holes

We have ignored the holes on the surface and the interior of entities in the definition of spatial extents. The presence of holes can be recovered from the use of the taxonomy and meronymic relations. First, consider holes in the interior. Since we encode that “a cytoplasm is inside a plasma membrane” but is not a part of it, we can infer that a membrane has space in its interior. Next, consider holes on the surface. Here, we encode that “a nucleus has a pore as a region”. Pores, channels, and the like are considered to be holes, and so we can recover the existence of holes on the surface.
4.9 Negative Information

In AURA, negative information can be encoded in three ways:
   a. Cardinality constraints – Does a human cell have 100 chromosomes? No, because it has exactly 46 chromosomes.
   b. Negation as failure – Does a prokaryotic cell (necessarily) have a nucleus? No, because we fail to find such an assertion.
   c. Disjointness and sufficient conditions – Can a prokaryotic cell (possibly) have a nucleus? No, because a cell that has a nucleus is a eukaryotic cell (sufficient condition), and eukaryotic and prokaryotic cells are disjoint.

In 2011, we considered expanding the types of negative information that could be encoded in a limited way, i.e., via the addition of negative slots. For example, we had the slot “prevent”, which is the negative counterpart of “enable”. Would it be useful to have negative counterparts of other slots? For example, does-not-have-part, not-object, and so forth? There are two associated questions:
   a. Does the use of negative slots add expressive power? To indicate absence, negative slots do not add expressive power. For example, we were able to encode the absence of a nucleus in a prokaryotic cell using disjointness and sufficient conditions. However, there is an increase in power when we wish to encode choice. For example, a membrane is a semi-permeable barrier, i.e., it is permeable to some things and impermeable to others. We have added the negative slots: is-impermeable-to, and is-unsaturated-with.

   b. Is it worthwhile to create the negative counterpart of all slots? Further analysis is needed.

4.10 States
States can be subdivided into states of entities and states of processes.

**States of Entities**
A state of an entity is an entity such that its slot values or property values change during a process, but it maintains identity across those changes. These include states in which there are unique biology names (e.g. blastula, gastrula), and states in which there are no unique biology names (e.g. cell that is younger, cell that is older).

States of entities are subclasses of entities that have the same structural attributes as their superclass, which facilitates comparison to their superclass and related states. For instance, an organism could be defined as having regions and a shape. Organism could have subclasses Zygote, Blastula, Gastrula, and so on. Figure 24 shows Blastula as having a specialized region and shape.
Figure 24: Blastula has inherited features that are specialized in the Blastula Concept.

State identity is tracked using *has-state* or its inverse *is-state-of*. For instance, in the overly simplistic Animal-Development in Figure 25, note that Animal (which is a type of Organism) has-state Zygote early in Animal-Development, then the state of the animal changes as the process continues. All states maintain the identity Animal, but each can have different attributes associated with them, defined on each state’s Concept, or here, if they were specific to Animal-Development.

Figure 25: Four states of an animal across an oversimplified Animal-Development process.
In the context of a process with several ordered events, the current state can only be the object of the immediate next event. For instance, it would be incorrect to encode that the Zygote was a raw material of Gastrulation; if Animal is being transformed into a new state, the only option is to use the state of the Animal at the end of the Cleavage event.

In the context of an entity changing parthood/possession, the has-state relation is useful for implying a removal or addition, without needing to overtly encode this. For instance, in a chemical reaction, Molecule A might have an Acetyl-Group region that is removed and then bound to another molecule, which results in Molecule B having this region, Acetyl-Group. Rather than encode events Add and Remove to make overt that the acetyl group being used is the one that was removed from Molecule A and subsequently added to form Molecule B, we could encode Chemical-Reaction with raw material Molecule A having region Acetyl-Group, and result Molecule B having Acetyl-Group. Adding a state-of-relation between the two Acetyl-Groups would indicate that they have the same identity.

**States of Processes**
Under development.

### 4.11 Biological Classification
Classification of entities is central to biology, and represents some interesting modeling challenges. In biology:

- Group names are associated with a taxonomic level (e.g. *coli* is a species, *Animalia* is a kingdom)
- There are features (or sometimes absence of a feature) associated with classification (e.g. shape, color, smell, anatomical structures, observed behaviors of a single organism, etc.)
- There is a scheme of relatedness between taxonomic levels and groups. For instance:
  - Kingdom “contains” several Phyla
  - Animalia and Plantae are sibling phyla
  - Eukarya domain is more related to Archaea domain than to Bacteria domain
  - Protista and Plantae diverged X years ago OR Protista and Plantae show X% similarity in Y gene sequence
- Groupings are disputed, human-designed organization schemes
  - Some organisms/groups fall under multiple classification groups
  - Some taxonomic groups are grouped into one scheme, and others are grouped into others
  - Linnaean Hierarchy includes Domains, Kingdoms, Phyla, etc.
  - and Super Groups are an alternative method of classification

We have resolved that there will be three distinct taxonomies in the knowledge base including:

1) Taxonomic Group taxonomy.
a. This includes Kingdom, Phylum, etc. This has not yet been implemented in the knowledge base.
b. Further planning needs to be done to determine if it will be broad and shallow (i.e. Kingdom and Phylum are siblings), or narrow and deep (i.e. Phylum is a subclass of Kingdom).
c. We may need a new relation such as *taxa-level* to indicate that a group of organisms is at a specific taxonomic level in a hierarchy (e.g. Animalia-Group – taxa-level – Kingdom).
d. Models can be created such as Linnaean-Classification to indicate how some taxonomic groups are grouped together. Linnaean-Classification would include domains and kingdoms, for instance, but not supergroups. We may need a new relation to show that Phylum is an "element" of Kingdom in this model.

2) Organism Group taxonomy.
   a. This would include groups with special names (e.g. Animalia is recognized as a Kingdom of animals), as well as groups whose names are identical to a single organism (e.g. Drosophila or *E. coli*). In the case of identical naming for the group and individual, then –Group will be added to the end of the name (e.g. Animalia (individual organism) – member –of- Animalia-Group).
   b. All entities in this taxonomy are mass-terms.
   c. Groups of this type are inclusive of kingdoms, species or populations; meaning that a group can be defined at a variety of scopes.
   d. A new relation *member-of* is needed to indicate that an individual organism belongs to a group.
   e. Groups that already exist in the knowledge base have not been edited or moved at this point, but should not be used as superclasses of individual organisms.

3) Organism taxonomy.
   a. This includes single organisms, such that each Concept is for an individual.
   b. The Concept would include required attributes for identification or classification of this organism type, including use of triggers and disjoints that could facilitate classification.
   c. Naming should reflect popular usage in *Campbell Biology*, meaning some classes can have plural names, even though the Concept is for a single individual. For instance, Archaea will not be renamed Archaeum, but could be renamed Archaea-Cell.
   d. The vast majority of the knowledge base already follows the convention of defining an organism as an individual. We will continue this practice.
4.12 Aggregates

4.12.1 Superclasses

We adopt the following guidelines:

- A concept has the superclass Aggregate if it is understood as a collection of two or more entities. See the following section for guidelines on identifying aggregates in the text.
- If a concept has the superclass Aggregate, it can have additional superclasses, such as Spatial-Entity, Region, Tangible-Entity, or Chemical-Entity. Additional superclasses should be added by checking whether the definition of those classes fits. For example, Molecular-Assembly can have the superclasses Aggregate, Chemical-Entity, and/or Molecule as appropriate.

4.12.2 Identifying Aggregates in the Text

During the encoding process, SMEs should look for terms such as “aggregate”, “assembly”, “collection”, or “group” to identify aggregates. This list of terms is by no means exhaustive. For example, given the sentence “A polymer is a long molecule consisting of many similar or identical building blocks linked by covalent bonds, much as a train consists of a chain of cars”, we may interpret “polymer” as a candidate aggregate.

When a sentence states that X is an aggregate of Y, the main decision is whether the entity is an Aggregate or whether the Aggregate refers to a collection of parts. (This is very similar to the issue of distinguishing between superclasses vs parts and subevents, discussed in Section Error! Reference source not found.) Consider the following sentences:

- The aggregate refers to a collection of parts – The sentence states that X is an aggregate of Y, but either X is a substance, or Y is much smaller than X (it is usually part of a part of X). In such cases, we use has-part and avoid the use of Aggregate. Consider the following examples:
  a. Some organisms, such as amoebas, consist of single cells, but others are multicellular aggregates of many specialized types of cells. (The sentence describes a Multicellular-Organism (X) as an aggregate of Cells (Y). But, cells are usually parts of parts of multicellular-organisms, and so we prefer Multicellular-Organism has-part Cell)
  b. Salts are often found in nature as crystals of various sizes and shapes, each an aggregate of vast numbers of cations and anions bonded by their electrical attraction and arranged in a three-dimensional lattice. (The sentence describes Salt as an aggregate of Anion and Cation (Y). But, salt is a substances, and so we prefer Salt has-part Cation)
- The entity is the aggregate – The sentence states that X is an aggregate of Y, X is not a substance, and Y is a direct part or element. In the following examples, we use Aggregate as a superclass

68
a. The ligand binding causes two receptor polypeptides to aggregate, forming a dimer (Dimer isa Aggregate)

b. The completed assembly of transcription factors and RNA polymerase bound to the promoter is called a transcription initiation complex. (Transcription-Initiation-Complex isa Aggregate)

c. A collection of identical or very similar genes is called a multigene family. (Multigene-Family isa Aggregate)

Unclear cases should be discussed in the call, and decisions taken on a case-by-case basis.