MIRIAM Resources: next steps

Camille Laibe

COMBINE, 8-10 October 2010, Edinburgh, UK

EMBL-EBI

- **Introduction**

- Potential issue

- Possible extensions

- Work in progress
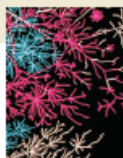
computational BIOLOGY

PERSPECTIVE

Minimum information requested in the annotation of biochemical models (MIRIAM)

Nicolas Le Novère[1,15], Andrew Finney[2,15], Michael Hucka[3], Upinder S Bhalla[4], Fabien Campagne[5], Julio Collado-Vides[6], Edmund J Crampin[7], Matt Halstead[7], Edda Klipp[8], Pedro Mendes[9], Poul Nielsen[7], Herbert Sauro[10], Bruce Shapiro[11], Jacky L Snoep[12], Hugh D Spence[13] & Barry L Wanner[14]

Most of the published quantitative models in biology are lost for the community because they are either not made available or they are insufficiently characterized to allow them to be reused. The lack of a standard description format, lack of stringent reviewing and authors' carelessness are the main causes for incomplete model descriptions. With today's increased interest in detailed biochemical models, it is necessary to define a minimum quality standard for the encoding of those models. We propose a set of rules for curating quantitative models of biological systems. These rules define procedures for encoding and annotating models represented in machine-readable form. We believe their application will enable users to (i) have confidence that curated models are an accurate reflection of their associated reference descriptions, (ii) search collections of curated models with precision, (iii) quickly identify the biological phenomena that a given curated model or model constituent represents and (iv) facilitate model reuse and composition into large subcellular models.

During the genomic era we have witnessed a vast increase in availability of large amounts of quantitative data. This is motivating a shift in the focus of molecular and cellular research from qualitative descriptions of biochemical interactions towards the quantification of such interactions and their dynamics. One of the tenets of systems biology is the use of quantitative models (see Box 1 for definitions) as a mechanism for capturing precise hypotheses and making predictions[1,2]. Many specialized models exist that attempt to explain aspects of the cellular machinery. However, as has happened with other types of biological information, such as sequences, macromolecular structures or

Box 1 Glossary

Some terms are used in a very specific way throughout the article. We provide here a precise definition of each one.

Quantitative biochemical model. A formal model of a biological system, based on the mathematical description of its molecular and cellular components, and the interactions between those components.

Encoded model. A mathematical model written in a formal machine-readable language, such that it can be systematically parsed and employed by simulation and analysis software without further human translation.

MIRIAM-compliant model. A model that passes all the tests and fulfills all the conditions listed in MIRIAM.

Reference description. A unique document that describes, or references the description of, the model, the structure of the model, the numerical values necessary to instantiate a simulation from the model, or to perform a mathematical analysis of the model, and the results one expects from such a simulation or analysis.

Curation process. The process by which the compliance of an encoded model with MIRIAM is achieved and/or verified. The curation process may encompass some or all of the following tasks: encoding of the model, verification of the reference correspondence and annotation of the model.

Reference correspondence. The fact that the structure of a model and the results of a simulation or an analysis match the information present in the reference description.

[1]European Bioinformatics Institute, Hinxton, CB10 1SD, UK. [2]Physiomics PLC, Magdalen Centre, Oxford Science Park, Oxford, OX4 4GA, K. [3]Control and Dynamical Systems, California Institute of Technology, Pasadena, California 91125, USA. [4]National Centre for Biological Sciences, TIFR, UAS-GKVK Campus, Bangalore 560065, India. [5]Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York 10021, USA. [6]Center for Genomic Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n, Cuernavaca, Morelos, 62100, Mexico. [7]Bioengineering Institute and Department of Engineering Science, The University of Auckland, Private Bag 92019, Auckland, New Zealand. [8]Max-Planck Institute for Molecular Genetics, Berlin Center for Genome based Bioinformatics (BCB), Ihnestr. 73, 14195 Berlin, Germany. [9]Virginia Bioinformatics Institute, Virginia Tech, Washington St., Blacksburg, Virginia 24061-0477, USA. [10]Keck Graduate Institute, 535 Watson Drive, Claremont, California 91711, USA. [11]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. [12]Triple-J Group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa. [13]Department of Scientific Computing & Mathematical Modeling, Gummels Wood Road, Stevenage, Herts, SG1 2NY, UK. [14]Purdue University, Department of Biological Sciences, Lilly Hall of Life Sciences, 915 W. State Street, West Lafayette, Indiana 47907-2054, USA. [15]These authors have contributed equally to the work. Correspondence should be addressed to N.L.N. (e-mail: lenov@ebi.ac.uk).

Published online 6 December 2005; doi:10.1038/nbt1156

- proposed guidelines for curation and annotation of quantitative models

- about encoding and annotation

- applicable to any structured model format

cf. Nicolas Le Novère _et al._ **Minimum Information Requested in the Annotation of biochemical Models (MIRIAM)**. _Nature Biotechnology_, 2005

http://biomodels.net/miriam/

EMBL-EBI

Models **must**:

- be encoded in a public machine-readable format

- be clearly linked to a single publication

- reflect the structure of the biological processes described in the reference paper (list of reactions, ...)

- be instantiable in a simulation (possess initial conditions, ...)

- be able to reproduce the results given in the reference paper

- contain creator's contact details

- annotated: **must unambiguously identify each model constituent**

EMBL-EBI

Models **must**:

- be encoded in a public machine-readable format

- be clearly linked to a single publication

- reflect the structure of the biological processes described in the reference paper (list of reactions, …)

- be instantiable in a simulation (possess initial conditions, …)

- be able to reproduce the results given in the reference paper

- contain creator's contact details

- annotated: **must unambiguously identify each model constituent**

Essential for data identification and **semantics**:

- Understanding

- Search

- Reuse

- Comparison

- Integration

- …

EMBL-EBI

Essential for data identification and **semantics**:

- Understanding

- Search

- Reuse

- Comparison

- Integration

- …

→ **True for any kind of data, not only models!**

- **Unique and unambiguous**

  an identifier must never be assigned to two different objects

- **Perennial**

  the identifier is constant and its lifetime is permanent

- **Standards compliant**

  must conform on existing *standards*, such as URI

- **Resolvable**

  identifiers must be able to be transformed into locations of online resources storing the object or information about the object

- **Free of use**

  everybody should be able to use and create identifiers, freely and at no cost

EMBL-EBI

Data type

Dataset Identifier

Not a URL, not a "Web-address"!

Format depends on the resource identified by the data type

Human calmodulin: P62158 in UniProt ➡ urn:miriam:uniprot:P62158

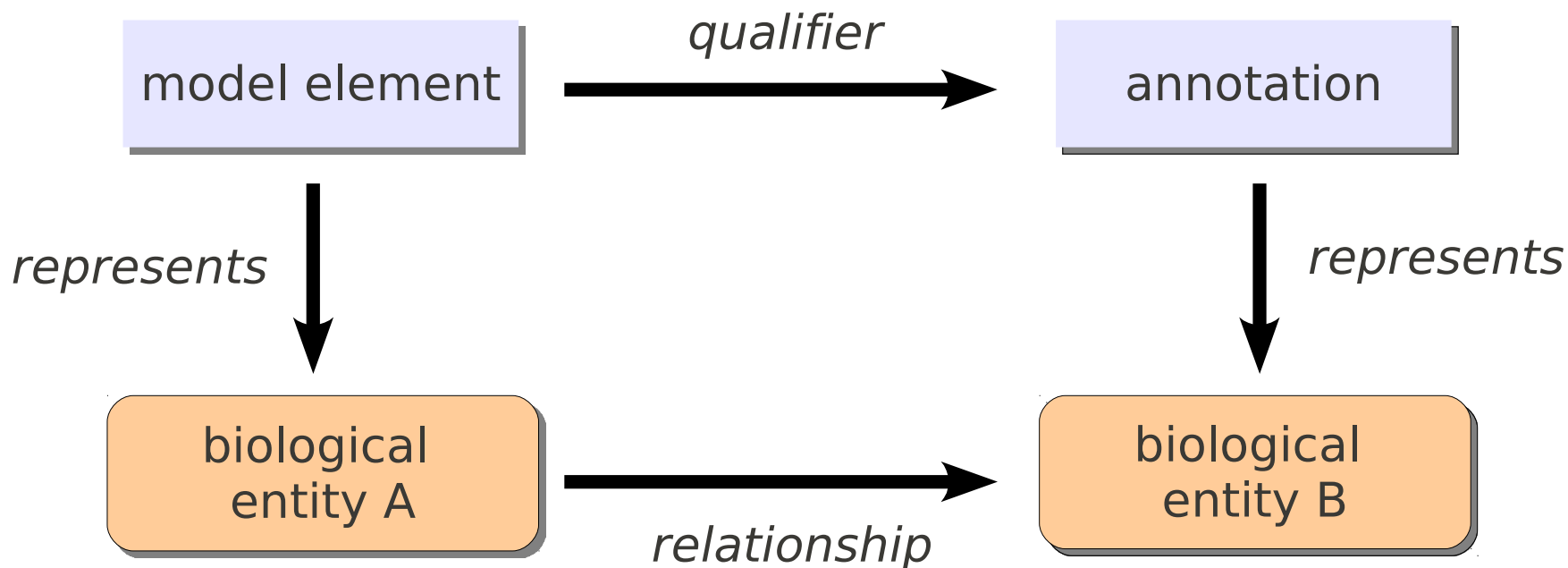Alcohol dehydrogenase: 1.1.1.1 in EC code ➡ urn:miriam:ec-code:1.1.1.1

Activation of MAPKK activity: GO:0000186 in Gene Ontology
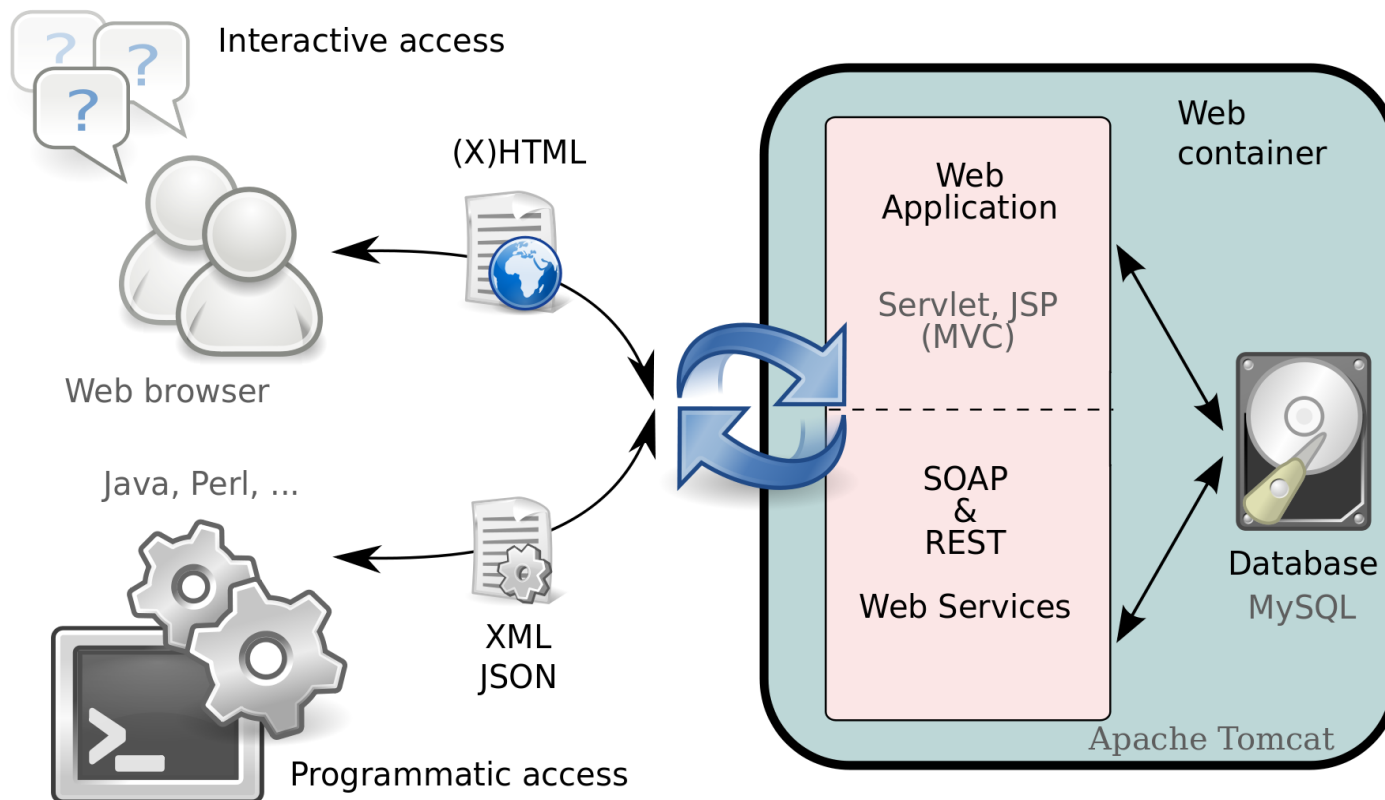
➡ urn:miriam:obo.go:GO%3A0000186

EMBL-EBI

model element → *qualifier* → annotation

represents

represents

biological entity A → *relationship* → biological entity B

- bqmodel:is

- bqmodel:isDerivedFrom

- bqmodel:isDescribedBy

- bqbiol:hasPart

- bqbiol:hasProperty

- bqbiol:isPartOf

- ...

http://biomodels.net/qualifiers/

Generation and resolving of MIRIAM URNs

Camille Laibe and Nicolas Le Novère.
MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology.
BMC Systems Biology, 2007

http://www.ebi.ac.uk/miriam/

- Introduction

- **Potential issue**

- Possible extensions

- Work in progress

- Current situation:

  Activation of MAPKK activity: GO:0000186 in Gene Ontology

  `urn:miriam:obo.go:GO%3A0000186`

- Issues:

  - need to encode ':' (not only in the context of MIRIAM URNs!)
  - duplication of the ontology identification
  - lots of complaints/remarks received...

- Possible solution:

  - remove the OBO namespace from the dataset identifier part

    `urn:miriam:obo.go:0000186`

- Introduction

- Potential issue

- **Possible extensions**

- Work in progress

EMBL-EBI

- Identification of a specific version of an entity:

  `urn:miriam:biomodels.db:BIOMD0000000008`

  - from 18th release

  - version from 30th September 2010

  - …

- Possible solutions:

  - Data provider issue: new identifier per revision

    `urn:miriam:biomodels.db:BIOMD0000000008_2`

  - Updated URN Scheme

    `urn:miriam:biomodels.db:BIOMD0000000008:2`

  - …

- Identification of an entity within an entity:

  `urn:miriam:`**`biomodels.db`**`:BIOMD0000000008`

  - species "protease" from this model
  - reaction "desinhibition of cyclin" from this model

- Possible solution:

  - Update URN Scheme:

    `urn:miriam:`**`biomodels.db`**`:BIOMD0000000008:`**`_202906`**

  - ...

- Introduction

- Potential issue

- Possible extensions

- **Work in progress**

EMBL-EBI

Identification of entities provided by data types which cannot currently be added to MIRIAM Resources

- Why?

  - needed by projects which receive data already (partially) annotated

- Example:

  - CAS (Chemical Abstracts Service) → not free

- Possible solution:

  - 2nd branch, with partial support (URN generation, but no other services provided, like resolving)

EMBL-EBI

- **Open access**

Anybody can access any public data without restriction (no commercial licence, no login page, …)

- **Atomicity**

The granularity of the data distributed has to be appropriately selected (a database of "reactions" distributes reactions and not pathways) and consistent (e.g. classes or instances but not classes *and* instances)

- **Identifier**

An atomic data is associated to a unique and perennial identifier

- **Community recognition**

The resource has to be "recognised" by the corresponding experimental community, be reasonably supported, …

Identify entities provided by data types which cannot currently be added to MIRIAM Resources

- Why?

  - needed by projects which receive data already (partially) annotated

- Example:

  - CAS (Chemical Abstracts Service)

- Possible solution:

  - 2nd branch, with partial support (URN generation, but no other services provided, like resolving)

EMBL-EBI

Resources are currently associated with one physical location (URL)

- Storage of multiple URLs per resource, with indication of the returned format:

  - `http://www.uniprot.org/uniprot/P12345` → TEXT/HTML

  - `http://www.uniprot.org/uniprot/P12345.txt` → TEXT/PLAIN

  - `http://www.uniprot.org/uniprot/P12345.xml` → TEXT/XML

  - `http://www.uniprot.org/uniprot/P12345.rdf` → RDF/XML

- Direct access to archives (whole datasets)

- Cross references to external providers when relevant:

  - BioCatalogue, for Web Services records

  - …

Provide support for any changes that might occur to the data MIRIAM
Resources provide → backward compatibility

- REST Web Services

    - currently in *beta*

- Disconnected and standalone library

    - no query over the Web for every single request

- Updated XML export

- ...

EMBL-EBI

- The community of computational systems biology
  for the development of MIRIAM and the implementation of MIRIAM support
- Data providers
  who replied, discussed and even complied with MIRIAM rules

    - Nicolas Le Novère
    - Nick Juty
    - Camille Laibe

EMBL

bbsrc
biotechnology and biological
sciences research council

NIGMS

ELIXIR

camille.laibe@ebi.ac.uk

- Potential issue:

    - OBO namespace in MIRIAM URNs

- Possible extensions of MIRIAM URN Scheme:

    - Identification of a specific revision of an entity

    - Identification of an entity within an entity

- Work in progress:

    - Creation of a set of data types with partial support only

    - Storage of multiple URLs per resource (with their returned format)

    - Direct access to whole datasets

    - Cross references to external providers