Metamodelling architectures for complex data integration in systems biology

Marie-Noelle Terrasse

LE2I University of Burgundy-CNRS UMR 5158,
UFR Sciences et Techniques, Bât. Sciences de l'Ingénieur,
Faculté Mirande, 21000 Dijon, France
E-mail: marie-noelle.terrasse@u-bourgogne.fr

Magali Roux*

Laboratory of Computer Sciences,
University Pierre and Marie Curie-CNRS UMR 7606,
104 avenue du Président Kennedy 75016 Paris, France
E-mail: magali.roux@lip6.fr
*Corresponding author

Abstract: Systems biology aims at deciphering the functioning of biological systems on the basis of the knowledge of their molecular components and the relations between such components. To address the issues involved, high-throughput technologies are used. Taking advantage of the standards that are being currently developed to achieve consensual representations of technological domains, we present a metamodelling architecture based on these standards. The proposed architecture organises standard-specific metamodels and models into a single hierarchy. Each metamodel describes a consensus that is shared by several models of applications. A metamodel construct for description of faceted element is proposed together with this architecture.

Keywords: metamodelling architecture; integration; complex data; systems biology; biological standards; FuGE; functional genomics experiment; MAGE-OM; PSI MI; proteomics standards initiative molecular interaction; MDE; model-driven engineering.

Reference to this paper should be made as follows: Terrasse, M.N. and Roux, M. (xxxx) 'Metamodelling architectures for complex data integration in systems biology', *Int. J. Biomedical Engineering and Technology*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Marie-Noelle Terrasse received her PhD Degree in Computer Science in 1991. She worked at the US National Institute of Standards and Technology. Currently, her research and teaching interests focus on metamodelling-based engineering of domain-specific frameworks involving complex and evolving application domains.

Magali Roux received her PhD Degree in Biochemistry and Molecular Biology in 1983 and currently serves as a research director at CNRS in Paris, France. Her research activity is mainly on knowledge organisation and conceptual modelling in systems biology. She has published a number of papers on wet (Molecular Immunology and Cancer) and dry (Systemic Modelling) biology.

1 Introduction

Biology in the post-genomics era uses high-throughput methodologies named omics to study the abundance and variety of genes (genomics), gene transcripts (transcriptomics), protein expression (proteomics), metabolic flux (metabolomics), etc. These methodologies use various formalisms to represent biology data, together with multiple ontologies to annotate them, thus leading to complex interdependencies among data. Portals and warehouses are being developed for accessing data from existing ontologies and databases. Portals directly access and use remote source data, while warehouses need to Extract, Transform, and Load (ETL) data in proprietary depositories. Warehouses, thus, require frequent synchronisation. Nevertheless, this dual variance on the final location of data is rather simplistic because it is not only data that are changed but also their representations (i.e., data models). In fact, as knowledge is augmented, data models are modified (e.g., relations may be added or removed between entities; concepts may be added, removed or split). Due to data complexity and intricacy, such model variations could result into semantic inconsistencies of retrieved data. In this respect, a central issue in data management for biology and medicine deals not only with data updating but also with model evolution.

Model-Driven Engineering (MDE) is an approach developed for complex information system design that provides methods for dealing with model and data intricacies. *An important aspect of MDE is the emphasis it puts on bridges between technological spaces and on the integration of bodies of knowledge developed by different scientific communities* (Favre, 2004) as it is the case in biology, particularly with omics methodologies. In this paper, we propose to adapt MDE to the problem of data integration in biology. We take advantage of data models and schemas developed for standard designs in omics sciences to build metamodelling architectures. Standards provide dedicated frameworks for describing experiments carried out in certain omics technological domains with the purpose of capturing biology data and annotating them; these frameworks propose domain-dedicated entity-relation schemas or UML class diagrams that can be considered as ontologies. They include, in particular, the MicroArray and Gene Expression (MAGE) model for transcriptomics (Brazma et al., 2001), and the Proteomics Standards Initiative Molecular Interaction (PSI MI) model for proteomics (Hermjakob et al., 2004). Using such consensual representations, tools can be developed to implement the standards, e.g., the ArrayExpress database (Parkinson et al., 2007) constitutes a common repository based on the MAGE model.[1]

Possible approaches for linking models from different omics technological domains were suggested in order to control data intricacy (Xirasagar et al., 2004). Nevertheless, this resulted in very large models that are difficult to handle. More recently, the Functional Genomics Experiment (FuGE) standard (Jones et al., 2006) focused on concepts shared among omics models. The purpose of FuGE is to offer a high-level integration model to be used as an umbrella for model design and standardisation. FuGE proposes major building blocks for constructing the core of a domain specific language for description of integrated biological data. In spite of these valuable efforts, the FuGE initiative lacks strong theoretical bases such as those that were developed in the field of MDE (Favre, 2004). As was previously suggested, convergence between the need for integration of complex biological data and metamodelling architectures introduced in

the field of MDE (Roux-Rouquié and Schuch da Rosa, 2006) could be helpful for ensuring consistency and traceability of quality in model design and complex data integration (Terrasse et al., 2006).

2 Related work

Two major classifications of integrated systems have been proposed, namely centralised versus decentralised architectures, data-driven versus model-driven systems. Proposed solutions range from data warehouses to federated databases and peer-to-peer systems. Peer-to-peer systems do not feature global schemas of stored data since they use local data mappings. Each database of a federation uses its own structure of information while the federated system as a whole uses an integrated schema that can be more or less accurate. Warehouses use mappings in order to extract, transform and load data from remote sources. Various heterogeneities, thus, may need to be addressed (Bouquet et al., 2004), such as syntactic heterogeneity due to representation formats, terminological heterogeneity due to naming conventions, conceptual heterogeneity due to coverage, granularity, perspective of sources, and pragmatic heterogeneity due to user interpretations. Mappings are of major importance for addressing these problems. The mappings are classified (Shvaiko and Euzenat, 2005) according to the inputs they accept (e.g., languages, levels of abstraction), the processes they use (e.g., exact or approximate classifications, syntactic or semantical processing, possible use of external resources), and the output they produce (e.g., one-to-one or many-to-many correspondences, use of equivalence relations or of multiple relations, definitions of confidence estimates). In order to address the complexity of ETL processes (Boussaid et al., 2007) and the need to facilitate scaling up of already used algorithms (Euzenat et al. 2007), composite approaches are being developed (Do and Rahm, 2002).

Several methods that can carry out integration based on various levels of abstraction of a metamodelling architecture have been proposed (Bézivin et al., 2005; Gašević et al., 2007) that combine models, metamodels and ontologies (Bézivin et al., 2005; Pan and Horrocks, 2001). In biology, more than 60 ontologies and 900 databases (Galperin, 2007) are reporting daily increase in knowledge, and data integration, thus, poses a major challenge for systems biology. Various authors have discussed the adaptation of existing approaches to integration in complex domains towards systems biology (Gardner, 2005; Louie et al., 2006). Ontological engineering is widely used when a strong consensus exists in research communities (Brockmans et al., 2006; Smith, 2004). Data warehouses are generally preferred for integration of closely related domains (Bukhman and Skolnick, 2001) since they allow to manage consequences of the source databases' evolution. Federated approaches are often used for applications that use complex data whose modifications must be spread fast (Stein and Thierry-Mieg, 1998; Gopalacharyulu et al., 2005). More recent peer-to-peer approaches are not yet being used in biology (Kirsten and Rahm, 2006); nevertheless, important initiatives based on web services are under development at the European level (Labarga et al., 2007).

3 From standards to metamodels

In this section, we discuss the role of standards in the modelling of the biological domain. As the semantics of a complex domain can be hardly captured by a single model, we propose a model-driven methodology for integration of biological data.

3.1 Standards as reusable models

MDE emphasises reuse through domain specific languages (Mernik et al., 2005) and combinations of metamodels (e.g., representation and migration, refinement and abstraction, weaving) (Favre and Nguyen, 2004). Reuse is the key concept for application domain descriptions. MDE expresses such descriptions at the metamodel level. Yet, building a model for a complex application directly from a domain metamodel requires a huge amount of work. Generic models can help to manage the amount of contextual knowledge (i.e., the knowledge that is specific to a given application).

Various technological spaces have been developed to address biology's complexity. Large communities of researchers have formed around each of these technological spaces. These efforts have led to a specific consensus in each of these spaces. Various representations relevant to technological tools and databases rely on these consensuses. Growth of knowledge in post-genomic biology and systems biology requires integration of data from all of these technological spaces. Biologists need methods and tools that offer an integrated view of the consensus related to technological spaces. Within a given consensus it is also necessary to manage variations due to various consensus-compliant tools and databases. Rather than building a single structure for representing integrated data, it is better to build architectures of representations. Resulting integration architectures allow taking into account knowledge's complexity and its extent under the established consensus. In order to integrate data on the basis of the already established consensus, such as ontologies and standards that have been defined in biology (e.g., MAGE, PSI MI), it is necessary to base such an integration architecture on models and ontologies that are widely used and recognised within the biological community.

In the above perspective on the biological domain, standards are described as structured models. For example, the MAGE-Object Model (MAGE-OM) standard is organised into 17 packages, 132 classes and 223 associations that allow defining all types of DNA arrays. The packages are related among themselves, e.g., the *BioMaterial* package is related to *BioAssay*, *BioEvent* and *Array* packages. Similarly, the PSI MI standard uses the *Entry* class as the top element of its description of molecular interaction. This class has an *InteractionList* element as its mandatory child. Among standards currently under discussion in the biology community we mention the Cellular Assay Object Model, the PSI MOD ontology for protein modifications, and the PSI GEL for gel electrophoresis.[2] All these standards are models since they describe knowledge without specifying their actual implementation, yet they define procedures for verifying compliance of implementation tools. For example, MAGE-stk (Spellman et al., 2002) is an open-source toolkit that implements the MAGE model, whereas BIND and IntAct are databases compliant with the PSI MI model. In the view of MDE, standards are reusable models in the sense that compliance with standards is controlled by established guidelines that allow variations and extensions of each standard. For example, the Gene Expression Omnibus (GEO) database (Barrett et al., 2007) accepts

some, but not all, of the MAGE features and takes into account only selected elements from the MAGE *ProtocolApplication* package.

3.2 FuGE as a model of models

Standards generally go beyond mere reusable models. Recently, in order to cope with the newly created standards for data from high-throughput biological experiments, the FuGE model was proposed (Jones et al., 2006, 2007). FuGE consists of two parts, namely a *FuGE Common* part (with packages *base, audit, description, ontology, protocol, reference,* and *measurement*) and a *FuGE Bio* part (with packages *conceptual molecule, data, investigation, and material*) that are used in many other standards (MAGE, PSI MI, PSI SP, etc.). FuGE indeed possesses all of the features of metamodels, i.e., models of models and those of languages, for comprehensive and accurate description of models as it is the case for the Meta Object Facility (MOF)[3] used to describe conceptual modelling languages.

In this respect, FuGE could be defined as:

- A set *Cu* of elements that are building blocks for description of standards (extensions of a standard are described as combinations of such building blocks).
- A set *Cv* of different views that correspond to the current and future standards.
- A set *Ca* of rules and constraints developed to ensure that extensions are well-formed.

For the demonstration purposes we will limit our MDE approach to the existing artefacts and goals (i.e., developing new standards for intensive biology). A more sophisticated architecture can be constructed for integration of even more complex data.

4 A metamodelling architecture based on biological standards

Most complex domains contain various families of applications. Such families are related to different models or even to different metamodels. A complex domain can thus be hardly described by a single metamodel attached to models. Such domains need to be described by several metamodels and models. Some of these metamodels describe fundamental domain features (e.g., measurement features), while other metamodels describe modelling constructs dedicated to the technical integration of domain knowledge (e.g., reduction of modelling-style effects). We organise metamodels and models into a metamodelling architecture such that:

- each metamodel describes what is shared among its subordinated metamodels and models
- each model describes what is shared among its subordinated models and applications.

We also use reusable models to describe families of applications. Such reusable models are partial (or somehow imprecise models) that are used as a basis for building a specific model of an application. This way, applications belonging to the same family benefit from a consensus much wider than that described in their metamodel. A MDE-based process is based on the following hypothesis: a metamodel describes the stable building blocks of the domain knowledge. These building blocks are subjected to evolution to a lower degree than model-level blocks. Thus, most evolution steps are carried out without metamodel changes. Such a hypothesis has two major consequences: metamodel-driven data access preserves data semantics over evolution steps, and new metamodel-driven ETL processes can be proposed. This section outlines the main features of a metamodelling architecture that can be used as a basis for metamodel-driven ETL.

In many areas, and in particular in biology, standards offer general features needed for describing the consensual knowledge in a given context. Standards-compliant tools induce a particular way of using a given standard. In this paper we propose an example metamodelling architecture that encompasses both the FuGE standard and the family of applications developed according to the FuGe extension guidelines (Jones et al., 2007). We first present (Section 4.1) the metamodelling architecture itself, i.e., an organisation of metamodels, reusable models, and models that express the known extent of consensus among FuGE's users. We then discuss (Section 4.2) the need for an underlying metamodel-level construct that will allow modelling-style variations in the representation of the core concepts. We finally discuss the use of model transformations for implementation of such a construct.

4.1 Organising metamodels and models of biological domain features

The proposed architecture is organised in a hierarchy so that models of two given applications are linked to the same reusable model or metamodel that describes the consensus shared within the application domain. In our example, depicted in Figure 1, we study the application domain of FuGE-compliant applications that have been developed in accordance with FuGE's extension guidelines (Jones et al., 2007). The GEO (Barrett et al., 2007) and ArrayExpress (Parkinson et al., 2007) applications were developed according to the MAGE specifications. Five other applications were developed according to the PSI MI specifications.

The most general domain-related metamodel is the FuGE metamodel itself, which we denote by *mm*_FuGE. This metamodel describes the consensus shared by two FuGE-compliant applications. We define families of applications that share a consensus more precise than the FuGE general consensus. In our example, applications that recognise the FuGE's extension guidelines constitute such a family of applications. A reusable model, denoted by *rm*_FuGE-extension, describes what is shared by all applications in this family.

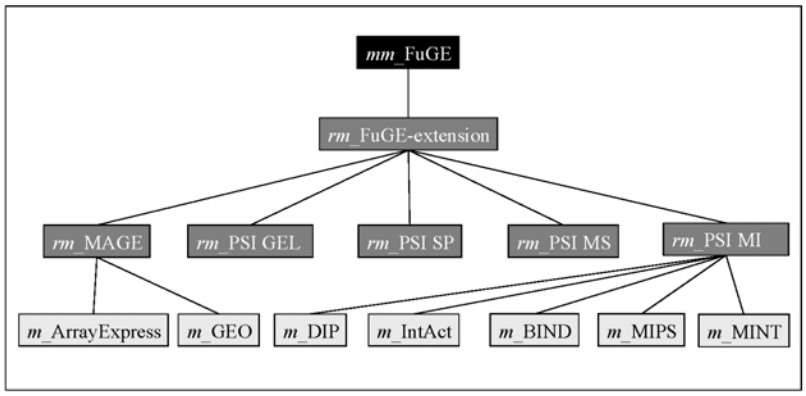
The MAGE specification defines a sub-family of the *rm*_FuGE-extension compliant applications. We denote the reusable model of this family by *rm*_MAGE. A model of ArrayExpress, denoted by *m*-ArrayExpress in our example, is derived directly from the *rm*_MAGE reusable model.

The group of specifications that are related to PSI MI defines another sub-family of the *rm*_FuGE-extension compliant applications. We denote the reusable model of this family by *rm*_PSI MI. In the metamodel architecture, models of DIP, IntAct, BIND, MIPS, and MINT are all derived directly from the rm_PSI MI reusable model. We denote these models by *m*_DIP, *m*_IntAct, *m*_BIND, *m*_MIPS, and *m*_MINT, respectively.

Other sub-families of the *rm*_FuGE-extension compliant standards are defined for gel-based experimental techniques, sample handling and processing, and mass spectrometry in proteomics. We denote their reusable models by *rm*_PSI GEL, *rm*_PSI SP, and *rm*_PSI MS, respectively.

The above metamodelling architecture facilitates interoperability by precisely defining what is shared and what is not shared by two applications. For example, the reusable model *rm*-PSI MI is shared by the IntAct and MIPS applications, while the *rm*_FuGE-extension reusable model is the only model shared by the IntAct and GEO applications. A more detailed example is given in Section 4.3.

Figure 1 A metamodelling architecture for FuGE (a metamodel in black, reusable models in dark grey, models in light grey)



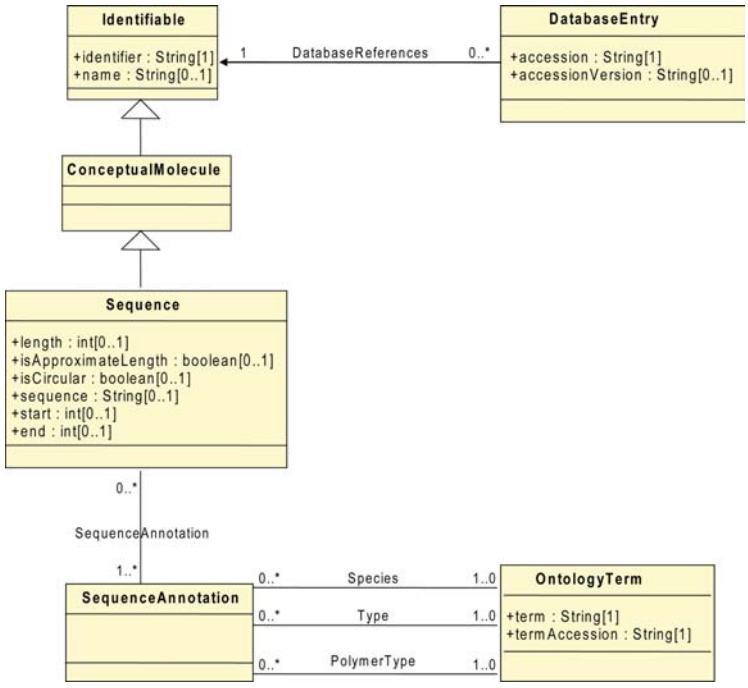
Since standards are constructed as general specifications, standards generally allow to model information in many different ways (corresponding to the modelling-style of users). For example, the *SequenceAnnotationSet* class in the FuGE model is a modelling artefact corresponding to the universe of the discourse of annotators, yet without correspondence to the universe of the discourse of experimental biology. In order to illustrate the modelling-style issue we discuss three models of a sequence: a *Sequence* model in FuGE (Jones et al., 2007) and two sequence models in FuGE-compliant specifications, namely, PSI MI and MAGE. When studying sequence descriptions in these three models, it is apparent that they use similar sets of features: name, identifier, length, species, etc., Nevertheless, two main aspects contribute to differences between models:

- optional features (that are not present in both models)
- variable features that are modelled differently (e.g., as an attribute or an association; different grouping of attributes, etc.).

The chosen example models are presented in Figures 2–4. Many sequence features are shared by all three models, and Table 1 summarises elements used to model a sequence in the chosen models:

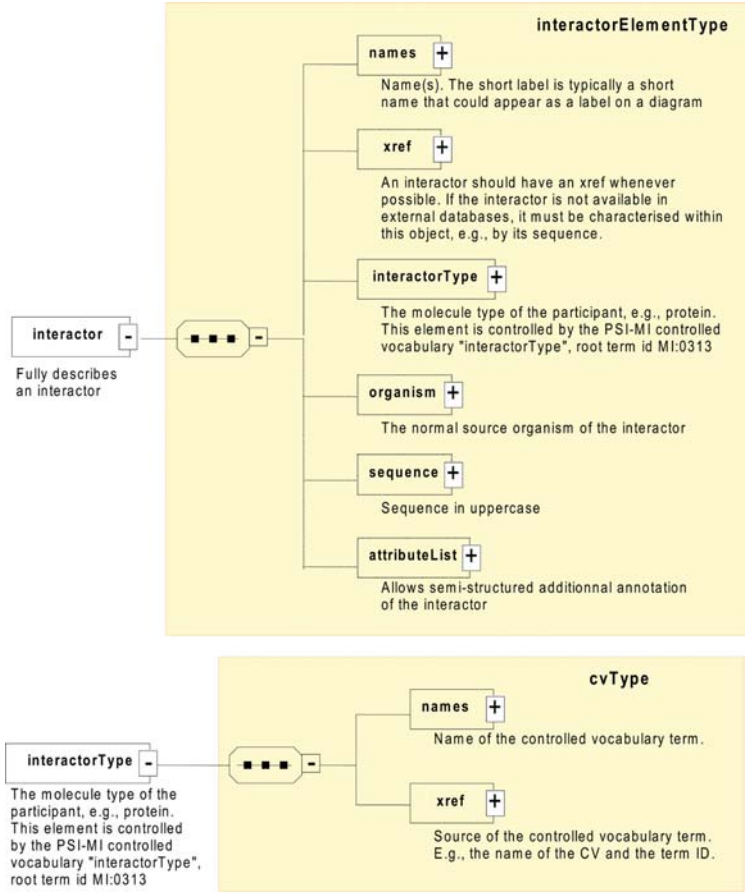
- The FuGE model (Figure 2) contains *ConceptualMolecule* as its main class. *ConceptualMolecule* is a specialisation of an *Identifiable* element. The *Identifiable* class is associated with *DatabaseEntry*(ies). The *Sequence* class is a specialisation of the *ConceptualMolecule* class and is associated with the *SequenceAnnotationSet*. The *SequenceAnnotationSet* itself is linked with *OntologyTerm*(s) by associations such as *Species*, *Types*, and *PolymerType.*
- The PSI MI XML schema (Figure 3(a)) is represented by a UML class diagram (Figure 3(b)) whose main component is the *interactor* class. This *interactor* class is associated (by an ordered aggregation) with classes *names*, *xref* (for database references), *sequence* (with few attributes), *organism*, *interactorType*, and *attributeList*. The *interactorType* class is associated (by an ordered aggregation) with classes *names* and *xref.* Two classes, *interactorElementType* and *cvType,* are needed to represent the corresponding grouping structures in the XML schema.
- The MAGE-OM model (Figure 4) has *BioSequence* as its central class. *BioSequence* is a specialisation of the *Identifiable* class, and it is associated with *DatabaseEntry*(ies) for bibliography references. The *BioSequence* class is associated with *OntologyTerm* (through the class *SequenceAnnotation*) by associations such as *OntologyEntries*, *PolymerType*, *Type*, *Species*.

Figure 2 Modelling of a sequence in the FuGE model (see online version for colours)

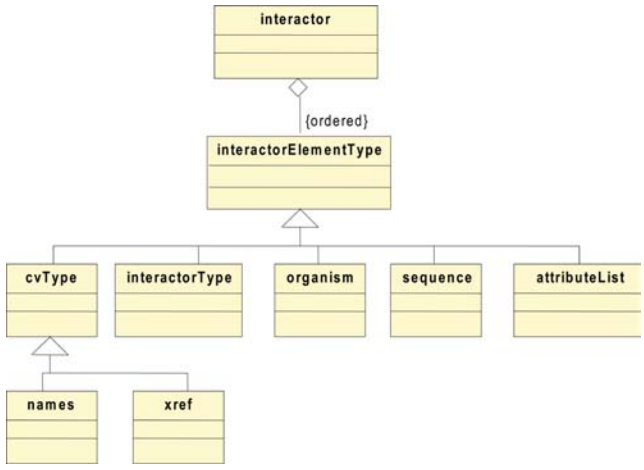


Source: Jones et al. (2006)

Figure 3 Modelling of a sequence in the PSI MI model (see online version for colours)



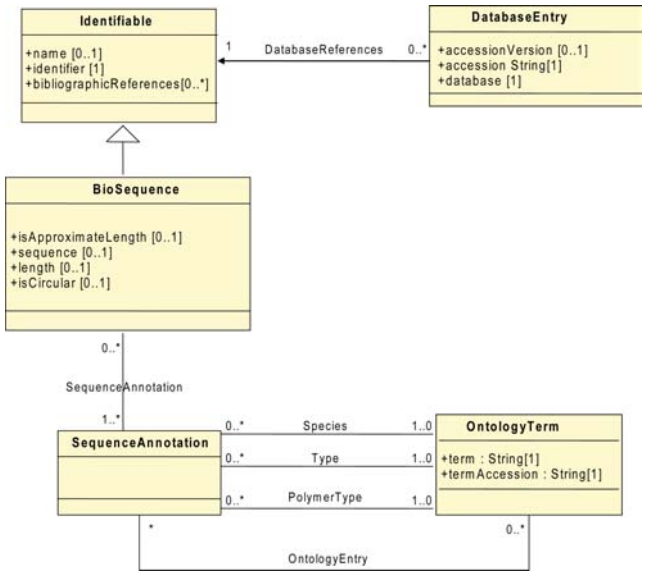
(a)



(b)

Source: Hermjakob et al. (2004)

Figure 4 Modelling of a sequence in MAGE-OM (see online version for colours)



Source: http://www.*omg*.org/cgi-bin/doc?formal/03-02-03

This comparison of sequence models shows that – even in the context of applying standards– variations are found in the modelled artefacts. Table 1 summarises the main variations found in these models.

Table 1 Comparison of sequence modelling in the FuGE, the PSI MI and the MAGE models

Elements of a sequence representation	*FuGe model*	*PSI MI model*	*MAGE model*
Length	Attribute of Sequence		Attribute of BioSequence
isApproximateLength	Attribute of Sequence		Attribute of BioSequence
identifier	Attribute of Identifiable		Attribute of Identifiable
name	Attribute of Identifiable	Class names	Attribute of Name
isCircular	Attribute of Sequence		Attribute of Sequence
sequence	Attribute of Sequence	Class sequence	Attribute of BioSequence
start, end	Attribute of Sequence		
Database reference	Association DatabaseReference	Class xref	Association SequenceDatabases
Species	Association species between classes OntologyTerm and SequenceAnnotationSet	Class Organism	Association species between classes OntologyEntry and BioSequence
types	Association types between classes OntologyTerm and SequenceAnnotationSet		Association polymerType between classes OntologyEntry and BioSequence
polymerType	Associations polymerType between classes OntologyTerm and SequenceAnnotationSet	Class InteractorType	Association polymerType between classes OntologyEntry and BioSequence

In order to account for these variations, we propose additional metamodel constructs for integration. Such constructs are presented in the following section.

4.2 Constructs for the integration metamodel

In order to improve application interoperability, we provide metamodel-level constructs that are able to reduce modelling-style and domain variations among various representations of the same entity (namely different groupings of features and differences between attributes and associations). Such a construct, called a *faceted element* (F_ELT) is intended for interoperability. It is not suitable for an initial modelling since this construct is relatively weak in terms of expressiveness (it reduces every element to a set of facets). Linking a F_ELT to an entity means that such an entity will be modelled differently in different views (e.g., a view for MAGE-OM, a view for PSI MI XML). A F_ELT is associated with *facet elements* (f_ELT) which represent the entity's attributes and relations. Models that use such a metamodel construct contain *faceted views* (F_VW) whose attributes and relations are also represented by *facet views* (f_VW). The F_ELT construct is thus built up from two parts that are linked together by associations and constraints (see Figure 5(a)):

- A F_ELT represents a core concept of the domain knowledge and is never used as a facet of another F_ELT.
- A F_ELT is defined in terms of f_ELTs. Facets may be optional or mandatory. Each facet represents an atomic component of the element's description (i.e., a facet is never further divided within models). f_ELTs are linked to the unique F_ELT they belong to by a defined element association (def_ELT, Figure 5(a)); def_ELT has the semantics of a *part_of* association.
- In a given model, a F_ELT appears as a F_VW to which there are attached as many f_VWs as necessary. f_VWs are linked to the F_VWs they belong to by a defined view association (def_VW, Figure 5(a)); def_VW has the semantics of a *part_of* association.
- The F_VW is linked to a f_VW by a *faceted association* (F_ASS, Figure 5(a)). Each f_VW is linked to a facet by a *facet association* (f_ASS, Figure 5(a)). All mandatory facets are required to be present in the f_VW; F_ASS and f_ASS have the semantics of a *is_a* association.

In our example, we use such constructs in the following way (see Figures 5(b) and 6):

1 We chose to describe a sequence model with F_ELTs at the upper-level of the metamodelling architecture, i.e., in the FuGE metamodel (Figure 5(b), top part). f_ELTs are added to this F_ELT sequence in order to represent attributes and associations related to sequence descriptions in FuGE. The precise extent of such a description is to be defined by biologists. For example, biologists need to decide whether the *conceptualMolecule* class belongs to the set of facets or whether it does not belong to this set (i.e., whether this class is of semantical interest in the application domain or whether it is just a mere modelling artefact).

Figure 5 Metamodel-level constructs for faceted elements (F_ELTs) (continuous thick line: F_ASS, dashed thick line: f-ASS , continuous thin line: def_ELT)

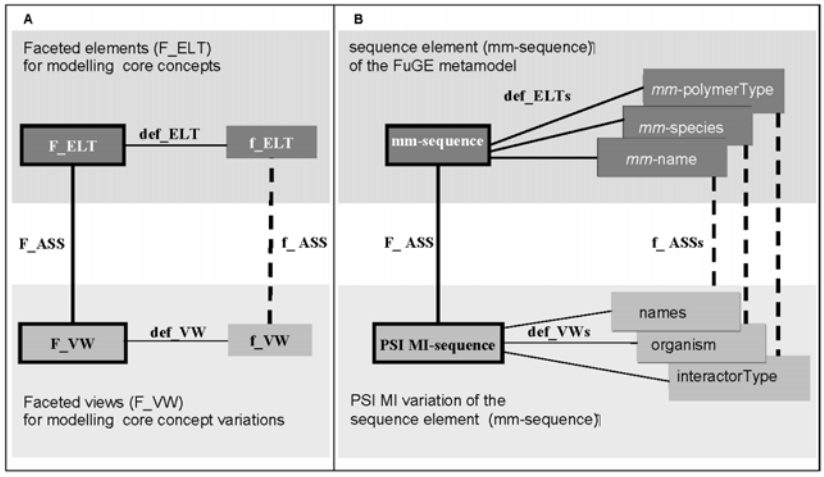
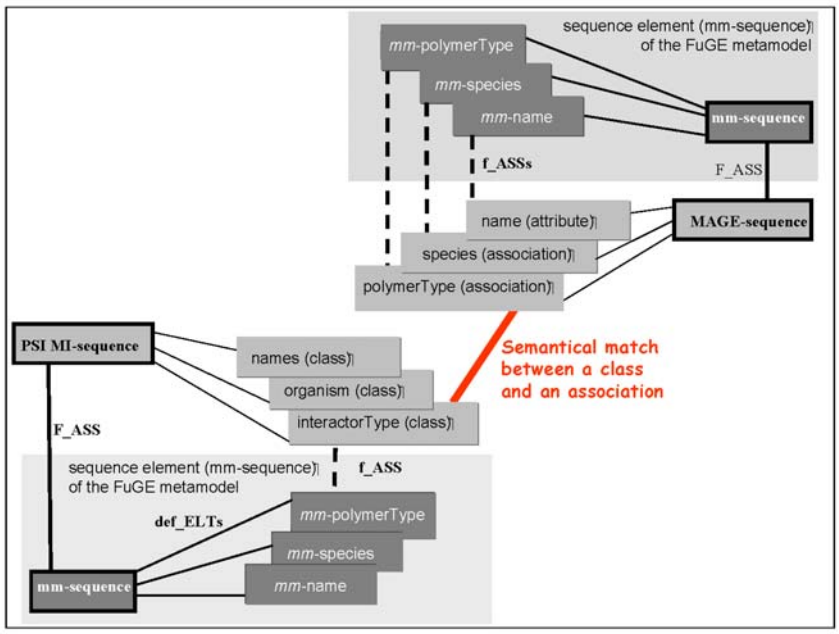


Figure 6 Using a faceted element (F_ELT) for description of PSI MI and MAGE-OM sequences (see Figure 5 for legend) (see online version for colours)



We define a F_ELT, denoted by *mm*-sequence, together with the f_ELTs listed below (two facets are specified as mandatory, namely *mm*-name and *mm*-polymerType*)*:

- *mm*-name and *mm*-identifier: facets of attributes of the generalised *Identifier* class
- *mm*-conceptualMolecule: a facet of the generalised *conceptualMolecule* class

- *mm*-databaseReference: a facet of an association of the generalised *Identifier* class
- *mm*-length, *mm*-isApproximateLength, *mm*-sequence, *mm*-isCircular, *mm*-start, *mm*-end: facets derived from attributes of the *Sequence* class
- *mm*-species, *mm*-types, *mm*-polymerType: facets derived from associations of the *SequenceAnnotationSet* class.

2 A F_VW of a sequence, denoted by PSI MI-sequence, is also defined for the PSI MI standard (Figure 5(b), in the lower half). The PSI MI-sequence view should be associated with f_VWs of the mandatory *mm*-name and *mm*-polymerType facets. We define classes *names* and *interactorType* as f_VWs for *mm*-name and *mm*-polymerType, respectively. The f_VW for the *organism* class in PSI MI model is associated with the *mm*-species f_ELT. The f_VW for the *sequence* class in the PSI MI model is associated with the *mm*-sequence f_ELT. In addition, the f_VW for the *interactor* class in the PSI MI model is associated with the *mm*-conceptualMolecule f_ELT. The f_VW for the *xref* class in the PSI-MI model is associated with the *mm*-databaseReference f_ELT (not shown in Figure 5(b)).

3 Analogously, we define a F_VW of a sequence for the MAGE-OM standard (Figure 6, in the upper half) We denote it as MAGE-sequence The MAGE-sequence view is associated with mandatory f_VWs that correspond to:

- the attribute name of the generalised *Identifiable* class
- the *polymerType* association.

The MAGE-OM associations *types* and *species* are associated (as f_VWs) with the f_ELTs *mm*-types and *mm*-species, respectively. Associations of f_VWs with certain features of sequences are straightforward since the MAGE-OM model is very close to the FuGE model: the *identifier* attribute of the *Identifiable* class; *length*, *isApproximateLength*, *isCircular*, *sequence* attributes of the *BioSequence* class; the *SequenceDatabases* association of the *BioSequence* class.

In Figure 6 we present an example semantical matching between the *polymerType* association of MAGE and the *interactorType* class of PSI MI. A model transformation of MAGE-compliant models into PSI MI compliant models should thus take into account such a matching. This matching is discussed in more details in the following section.

4.3 Model transformations

A model transformation turns a source model into a target model (Bézivin et al., 2005). Some of the constructs of the source model (e.g., classes, attributes, associations) are changed in order to produce building blocks of the target model. In this section, we first present an example transformation between models that relate to standards proposing similar models of a sequence (namely, a FuGE compliant model and a MAGE model). We then discuss the use of our facet stereotypes to control semantical accuracy of transformations between models that relate to standards proposing rather different models of the same biological concept.

Let us consider a FuGE compliant application, denoted by A, whose model *m-A* contains the model of sequences given in Figure 2. We wish to translate sequence-related data of the application A for the use by another application B whose model contains a sequence modelling derived from the MAGE model (see Figure 4). The following transformation steps are to be performed in order to transform m-A (or more precisely, the part of m-A related to sequences) into a MAGE compliant model:

1 Verification of the guarding condition: "classes *ConceptualMolecule*, *Sequence,* and *Identifiable* exist in the source model m-A".

2 Completion of m-A:

- If the *DatabaseEntry* class does not exist in m-A, then import it from the *Reference* package together with the *DatabaseReference* association
- If the *OntologyTerm* class does not exist in m-A, then import it from the *Ontology* package
- If the *SequenceAnnotationSet* class and the *SequenceAnnotation* association do not exist in m-A, then create them.

3 Transformation of the *Sequence* and *ConceptualModecule* classes:

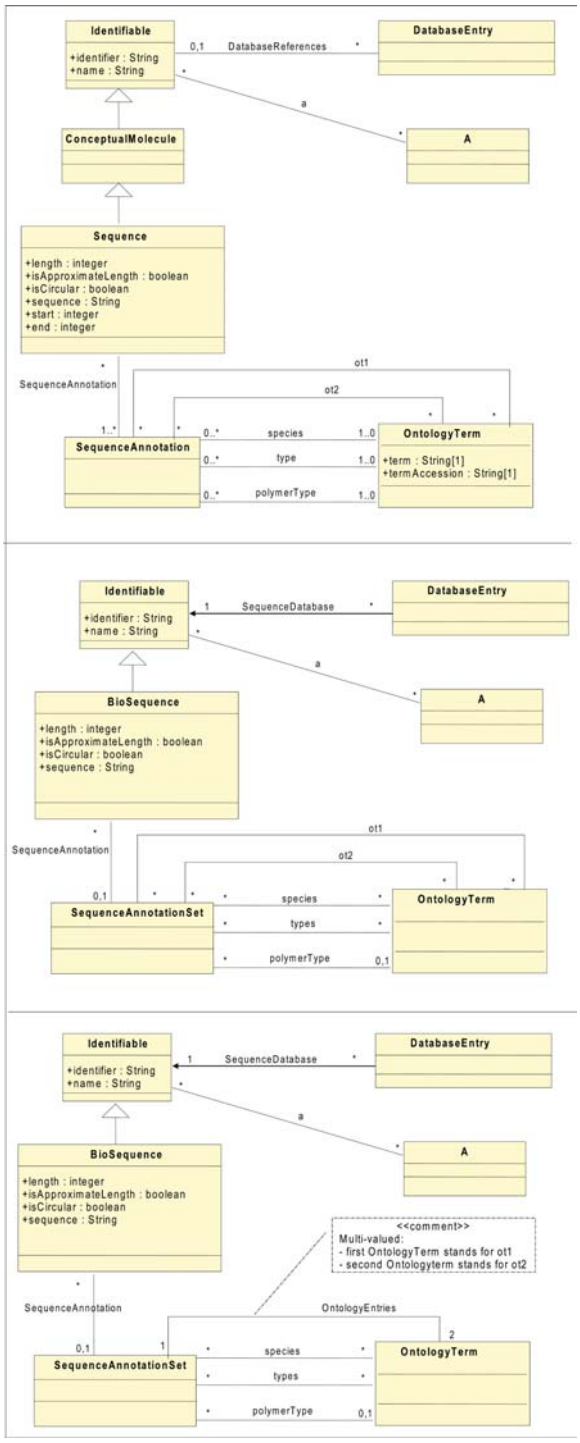
- Move all association-ends from the *ConceptualMolecule class* to the *Sequence* class. Association names, multiplicities and roles remain unchanged
- Delete the attributes *start*, and *end* from the *Sequence* class
- Insert a specialisation link from the *Sequence* class to the *Identifiable* class
- Delete the *ConceptualMolecule* class
- Rename the *Sequence* class to *BioSequence*.

4 Transformations of associations of the OntologyTerm class:

- Association *species*, *types*, and *polymerType* remain unchanged
- If additional associations exist between the *SequenceAnnotationSet* and *OntologyTerm* classes, such associations need to be gathered into a single new association, multivalued at the *OntologyTerm* end, and named *OntologyEntries*.

An example model transformation is given in Figure 7. This example illustrates our transformation's behaviour at two major variation points between the source and the target models. First, a class denoted by A is associated with the *ConceptualMolecule* class in the source model. This *ConceptualMolecule* does not exist in the target model. Second, two associations, denoted by *ot*1 and *ot*2, are defined between the *SequenceAnnotationSet* class and the *OntologyEntry* class (while the only possible association in the target model is *OntologyEntries*) in the source model. The example source model denoted by *m-AE* presents three differences when compared with the basic FuGE model: an association *a* between the *ConceptualMolecule* class and the *A* class, and two associations (named *ot1* and *ot2*) between the *SequenceAnnotationSet* and *OntologyEntry* classes. The top part of Figure 7 shows the model *m-AE* before transformation, the middle part presents the model immediately after the completion of step 3, and the bottom part presents the resulting model (after the completion of step 4).

Figure 7 An example model transformation. top: initial model m-AE (FuGE-compliant); middle: after completion of steps 1–3; bottom: transformed model (*MAGE-OM* compliant) (see online version for colours)



Beyond this introductory example, we intend to use our architecture so that facets can play a role of semantical guidelines for definition of elementary changes from which a model transformation is to be built. For example, a transformation between PSI MI and MAGE models should guarantee that the *polymer* association and the *species* association in a MAGE sequence model are transformed into an *interactor* class and an *organism* class in the PSI MI sequence model, respectively. Our facet stereotypes are, thus, to be complemented with rules that control model transformations. Such rules will make it mandatory for a model transformation to transform any f_VW of the source model into the f_VW of the target model that is linked (by f_ASS associations) to the same f_ELT. We are currently working on defining such rules in the general case.

5 Outlook

The MDE-based architecture of metamodels and models can be used to improve ETL processes, reuse and interoperability. An underlying interest of metamodelling architectures is to guide extraction, transformation processes in warehouses. First, metamodels and generic models that are shared by source databases can be used as filters for the extraction process since they define a domain consensus. Second, metamodels and generic models that are integrated in a warehouse metamodelling architecture can be used as guidelines for the transformation process. Third, abstraction levels of the metamodelling architecture can be used as guidelines for the synchronisation process. As long as a given metamodel or generic model are not modified in their source databases, their corresponding target metamodels and generic models can be used without synchronisation in the warehouse (thus allowing to restrict synchronisation to the data level).

Furthermore, metamodelling architectures can be used to design 'virtual and polymorphic' warehouses. A virtual warehouse contains a metamodelling architecture and the bindings of metamodels and models to actual data sources. For a given application, a sub-metamodelling architecture (containing only metamodels and models that are relevant to the application) can be defined. From this sub-architecture, the integrated metamodels and models of the application can be constructed by using model transformations. Queries are formulated for the warehouse depending on the integrated models and metamodels. Portals and web services are used to extract data from source databases and transform them according to the integrated metamodels and models.

The MDE approach allows incremental integration of knowledge, thus improving semantical consistency of a domain description. For example, data from MAGE-ML compliant applications (on gene expression) are used to identify genes involved in the transcription regulation; such data can be integrated with data from PSI MI applications on protein-protein interactions to infer networks in which the gene product can be involved. Such an approach is helpful in directing therapeutic targeting; otherwise integrating and comparing data from different experiments or different species are used to virtually increase the number of data samples under consideration. The architecture of standards, as well as the use of constructs for integration (e.g., facet/faceted), give certain clues for improving semantic consistency:

- *Constraint checking*. - The architecture allows reusing constraints checking. Certain rules need to be added to metamodels and models. Some of these rules are necessary to make the architecture's diagrams (metamodels or models) more precise. Other rules are defined in order to restrict standards' extensions (i.e., how to create new standards from existing ones). For example, a rule that requires an extension of the *Action* class of FuGE, if additional attributes are needed in this class, is formulated[4] "*Action SHOULD be extended if the Action requires additional attributes, for example to capture the role or function of the Action with respect to the parent Protocol*". In general, constraints should be verified, yet such verification may turn into a unmanageable task. By using the above metamodelling architecture, constraints that are already satisfied at the parent level and are unrelated to extended components do not need to be re-verified. In order to decide what needs to be verified, an approach similar to the one proposed in Cabot and Teniente (2007) can be applied.
- *Model transformations*. - The architecture allows reusing model transforms. For example, the ArrayExpress is compliant with MAGE, thus the transformation of ArrayExpress into a non-FuGE compliant model *m* should share most aspects of the transformation of MAGE into the model *m*. Such an incremental construction of transforms facilitates homogeneity of transforms and, thus, limits risks of semantical variations.

6 Conclusion

In this paper we described an organisation of biological standards and ontologies in metamodelling architectures which contain

- metamodels for description of general and widely recognised consensuses
- reusable models for description of specific consensuses
- models for description of applications.

We describe an example architecture for description of extensions of FuGE.

The faceted metamodelling construct allows to improve searching for items of knowledge since the needed pieces of information can be searched for without restrictions on specific representations. When searching for data related to a given facet view, we should look for resources that deal with the corresponding faceted view.

Faceted elements and model transforms are the key constructs of the proposed integration metamodel. Yet, in order to obtain better semantical accuracy, it is important to take into account all of the available information, e.g., association multiplicities, nested structures due to complex attributes of classes, etc. Most of this complementary information allows expressing constraints related to uses of faceted constructs and triggers in model transforms. Among main advantages of using such an architecture it is the reusability of constraints and model transforms that is to be emphasised:

- constraints can be expressed at various levels of the architecture, thus facilitating their verification
- model transforms can be defined incrementally, going from a reusable model to models derived from it.

Another main advantage of using such an architecture concerns improvements in interoperability: interoperating applications can be given a precise description of the consensus they share which each other, while at the same time avoiding problems due to variations of modelling styles in different applications (since the faceted element construct can be used).

In spite of efforts in standardisation, maintaining the semantical quality of integrated models is hard to assure. In order to achieve this semantical quality, we are currently exploring the use of the Basic Formal Ontology (BFO) as a semantical guideline for identifying consistent views of faceted elements (i.e., for defining facets, as well as for associating facet views with facets). BFO has been designed as an ontology applicable in all domains and consists of two hierarchies of concepts corresponding to continuants and occurrents (Grenon and Smith, 2004). Occurrents are entities that occur at a given time. Continuants are entities that have a longer existence and can change (in the sense that their properties vary). Three sub-concepts of continuants, namely substantial, property, and spatial region can be applied to facets we defined in our sequence example. A simple example is offered in the FuGE model where a *Sequence* is a substantial with properties such as *length* and *sequence string*. In the MAGE-OM model, a more slightly complex example is given with the *BioSequence* location that is described in terms of a spatial region (the *SeqFeatureLocation* and the *sub-region* association) and coordinates (*start* and *end* BFO properties). Using BFO may be of major interest in order to distinguish a processus (i.e., the execution of a procedure) and the static description of a procedure (i.e., a sequence of actions to execute, an execution environment, and guarding conditions). For example, in the *Protocol* package of FuGE the *protocol* class is a static description of a protocol which consists of elementary actions (*action* class). Executions of protocols and actions are modelled by two classes *protocolApplication* and *actionApplication*, respectively. Such a distinction between static and dynamic elements is essential and can be based on the BFO classification.

More generally, alignment with general ontologies is likely to provide good stability under diverging evolutions of biology sub-domains. Since biological data tend to be complex, major building blocks (e.g., representations of molecular sequences) generally depict several views which can differ from one model to another. Alignment of such views on concepts forming the core of a general ontology helps to guarantee their stability.

Acknowledgements

The authors thank George Becker for careful reading of the manuscript.

References

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) 'NCBI GEO: mining tens of millions of expression profiles – database and tools update', *Nucleic Acids Res*., Vol. 35, pp.D760–765.

Bézivin, J., Devedžić, V., Djurić, D., Favreau, J.M., Gašević, D. and Jouau, F. (2005) 'An M3-neutral infrastructure for bridging model engineering and ontology engineering', *Proc. INTEROP-ESA'05*, pp.159–171.

Bouquet, P., Euzenat, J., Franconi, L., Serafini, L., Stamou, G. and Tessaris, G. (2004) 'Specification of a common framework for characterizing alignment', *Knowledge Web Consortium*, EU Project IST-2004-507482.

Boussaid, O., Tanasescu, A., Bentayeb, F. and Darmont, J. (2007) 'Integration and dimensional modelling approaches for complex data warehousing', *Journal of Global Optimization*, Vol. 3, pp.571–591.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) 'Minimum information about a microarray experiment (MIAME–towards standards for microarray data', *Nat. Genet.*, Vol. 29, No. 4, pp.365–371.

Brockmans, S., Colomb, R.M., Kendall, E.F., Wallace, E.K., Welty, C. and Tong Xie, G. (2006) 'A model driven approach for building OWL DL and OWL FULL ontologies', *Proc. 5th Int. Semantic Web Conference* (*ISWC'06*), pp.187–200.

Bukhman,Y.V. and Skolnick, J. (2001) 'BioMolQuest: integrated database-based retrieval of protein structural and functional information', *Bioinformatics*, Vol. 17, No. 5, pp.468–478.

Cabot, J. and Teniente, E. (2007) 'Transformation techniques for OCL constraints', *Journal Science of Computer Programming*, Doi:10.1016/j.scico.2007.05.001, Vol. 68, No. 3, pp.179–195.

Do, H.H. and Rahm, E. (2002) 'COMA–a system for flexible combination of schema matching approaches', *Proc. 28th Very Large DataBases Conf. (VLDB'02)*, pp.610–621.

Euzenat, J., Petit, J.M. and Rousset, M.C. (2007) 'DECOR, passage l'échelle des techniques de découverte de correspondance', (Scaling up of techniques for correspondance discovery), *EGC Workshop*, Namur, Belgium, http://liris.cnrs.fr/publis/?id=2671

Favre, J.M. (2004) 'Towards a basic theory to model model driven engineering', *Workshop on Software Model Engineering, WISME'04*, Lisbon, Portugal.

Favre, J.M. and Nguyen, T. (2004) 'Towards a megamodel to model software evolution through transformations', *Proc. Workshop on Software Evolution through Transformations: Model-based vs. Implementation-level Solutions (SETra 2004), Electronic Notes in Theoretical Computer Science*, Vol. 127, No. 3, pp.59–74.

Galperin, M.Y. (2007) 'The molecular biology database collection', *Nucl. Acid. Research*, Vol. 35, pp.D3, D4.

Gardner, S.P. (2005) 'Ontologies and semantic data integration', *Drug Disc. Today*, Vol. 10, No. 14, pp.1001–1007.

Gašević, D., Djurić, D. and Devedžić, V. (2007) 'MDA-based automatic OWL ontology development', *International Journal on Software Tools for Technology Transfer*, Vol. 9, No. 2, pp.103–117.

Gopalacharyulu, P.V., Lindfors, E., Bounsaythip, C., Kivioja, T., Yetukuri, L., Hollmén, J. and Orešič. M. (2005) 'Data integration and visualization system for enabling conceptual biology', *Bioinformatics*, Vol. 21, Suppl. 1, pp.i177–i185.

Grenon, P. and Smith, B. (2004) 'SNAP and SPAN: towards dynamic spatial ontology', *Spatial Cognition and Computation*, Vol. 4, No. 1, pp.69–104.

Pan, J.Z. and Horrocks, I. (2001) 'Metamodeling architecture of web ontology languages', *Proc. Semantic Web Working Symposium (SWWS'01)*, pp.131–149.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. and Apweiler, R. (2004) 'The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data', *Nature Biotechnology*, Vol. 22, pp.177–183.

Jones, A.R, Pizarro, A, Spellman, P, and Miller, M (2006) 'FuGE: Functional Genomics Experiment object model', *FuGE Working Group OMICS Summer*, Vol. 10, No. 2, pp.179–184.

Jones, A.R., Miller, M., Aebersold, R., Apweiler, R., Ball, C.A., Brazma, A., Degreef, J., Hardy, N., Hermjakob, H., Hubbard, S.J., Hussey, P., Igra, M., Jenkins, H., Julian Jr., R.K., Laursen, K., Oliver, S.G., Paton, N.W., Sansone, S.A., Sarkans, U., Stoeckert Jr., C.J., Taylor, C.F., Whetzel, P.L., White, J.A., Spellman, P. and Pizarro, A. (2007) 'The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics', *Nat. Biotechnol.*, Vol. 10, pp.1127–1133.

Kirsten, T. and Rahm E. (2006) 'Biofuice: mapping-based data integration in bioinformatics', *Proc. 3rd Int. Workshop on Data Integration in the Life Sciences (DILS), LNCS 4075*, pp.124–135.

Labarga, A., Valentin, F., Anderson, M. and Lopez, R. (2007) 'Web services at the european bioinformatics institute', *Nucleic Acids Research*, Vol. 35, pp.W6–11.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. and Tarczy-Hornoch. P. (2006) 'Data integration and genomic medicine', *Journal of Biomedical Informatics*, Vol. 40, No. 1, pp.5–16.

Mernik, M., Heering, J. and Sloane, A.M. (2005) 'When and how to develop domain-specific languages', *ACM Computing Surveys*, Vol. 37, No. 4, pp.316–344.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U. and Brazma, A. (2007) 'ArrayExpress – a public database of microarray experiments and gene expression profiles', *Nucleic Acids Research*, Vol. 35, pp.1–4.

Roux-Rouquié, M. and Schuch da Rosa, D. (2006) 'Ten top reasons for systems biology to get into model driven engineering', Int. Conf. on Software Engineering. Proc. Int. Workshop on Global Integrated Management, Session: Metamodels and Semantics, Shanghai, China.

Shvaiko, P. and Euzenat, J. (2005) 'A survey of schema-based matching approaches', *Proc. Very Large DataBase Conf. (VLDB'05)*, pp.994–1005.

Smith, B. (2004) 'Beyond concepts: ontology as reality representation', in Varzi, A. and Vieu, L. (Eds.): *Proc. Third Int. Conf. on Formal Ontology and Information Systems (FOIS'04)*, pp.73–84.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert Jr., C.J. and Brazma, A. (2002) 'Design and implementation of microarray gene expression markup language (MAGE-ML)', *Genome Biol.*, Vol. 3, pp.research0046.1–0046.9.

Stein L.D. and Thierry-Mieg J. (1998) 'Scriptable access to the caenorhabditis elegans genome sequence and other ACeDB databases', *Genome Research*, Vol. 8, No. 12, pp.1308–1315.

Terrasse, M.N., Savonnet, M., Leclercq, E., Becker G. and Grison, T. (2006) 'Do we need metamodels and ontologies for engineering platforms ?', *Proc. 1st ICSE Int. Workshop on Global Integrated Model Management*, China, pp.21–27.

Xirasagar, S., Gustafson, S., Merrick, B.A., Tomer, K.B., Stasiewicz, S., Chan, D.D., Yost 3rd, K.J., Yates 3rd, J.R., Sumner, S., Xiao, N. and Waters, M.D. (2004) 'CEBS object model for systems biology data', *SysBio-OM*. *Bioinformatic*, Vol. 20, pp.2004–2015.

Notes

[1]http://www.omg.org/cgi-bin/doc?formal/03-02-03

[2]http://www.nature.com/nbt/consult/index.html

[3]OMG (2006) *Meta Object Facility Core Specification*, Version 2.0. URL http://www.omg.org/technology/documents/formal/MOF_Core.htm

[4]Guidelines for developing extensions on the FuGE Object Model (Draft guidelines based on the FuGE version 1 (candidate). A. Jones URL: http://fuge.sourceforge.net/