JISC

Technology & Standards Watch

An ontology methodology and

CISP

- the proposed Core Information about Scientific Papers

by

Larisa Soldatova and

Maria Liakata

Version	1.0
First published	December 2007
Publisher	JISC
Copyright owner	The University of Wales, Aberystwyth
Status of the document:	project report (intermediate)
Project:	ART – an ontology based tool for preparation of articles

-

Executive Summary

This report has two main goals:

- To introduce a new formalism for the description of scientific papers CISP (the Core Information about Scientific Papers);
- Attract more attention to ontologies as a valuable methodology for developing metadata.

Ontologies are a basic infrastructure for the Semantic Web. The idea of the Semantic Web is based on the possibility of using shared vocabularies for describing information resources. Ontologies provide unambiguous and machine-processable semantic metadata for Semantic Web applications. Ontologies are built on rigid theoretical foundations and can be used as proof of what concepts are essential for the description of a particular domain.

We demonstrate advantages of an ontology methodology for developing metadata by applying it to the analysis of the Dublin Core metadata (DC). An ontology approach allows detecting potential weaknesses in the representation of the DC terms. Such weaknesses include overlap in the semantic meaning between the terms, logically incoherent representation of temporal and spatial relations as well as incoherence in the representation of content. An ontology can also suggest improvements to the DC.

We used an ontology methodology to construct CISP metadata about the content of papers. It makes use of an ontology of experiments EXPO proposed at the University of Wales, Aberystwyth as a core ontology, and DOLCE (a Descriptive Ontology for Linguistic and Cognitive Engineering) developed at the Laboratory for Applied Ontology, the Institute of Cognitive Science and Technology, Italy as an upper level ontology. CISP is a defined set of leaf classes from these ontologies. It includes such key classes as <Goal of investigation>, <Object of investigation>, <Research method>, <Result>, <Conclusion>.

CISP can be used to generate abstracts and summaries of papers and also to facilitate storage and retrieval of information. CISP will constitute the basis for the ART tool. The latter is an authoring tool for the semantic annotation of papers stored in digital repositories. ART is intended for the semi-automatic annotation of data and metadata describing the scientific investigation represented in a research paper. ART will also be able to aid in the expression of research results directly in both a human and machine readable format, through the composition of text using ontology-based templates and stored typical key phrases. .

To find out more about ontology methodology refer to chapters 2 and 3 . To learn about the proposed CISP metadata you can start reading from chapter 4 onwards.

Status of the document: this is an intermediate project report about one of the outcomes of the ART project. We plan to submit a final version of the report about CISP by the end of the project (March, 2009).

Table of Contents

1. Introduction

Semantic Web technologies are focussed on using semantic metadata. Rich semantic representation can improve knowledge formalization and information retrieval. “Metadata can be defined literally as "data about data," but the term is normally understood to mean structured data about digital (and non-digital) resources that can be used to help support a wide range of operations”[1]. “Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use, or manage an information resource” [NISO 2004].

There are various sets of metadata: the Dublin Core Metadata[2], the Meta Content Framework (MCF)[3], Platform for Internet Content Selection (PICS)[4], RDF Site Summary (RSS)[5]. We will describe in this report the Dublin Core Metadata (DC) as one of the most popular and the most relevant set to our project. We will apply an ontology methodology for the analysis of the DC to demonstrate the advantages of our approach for the development of metadata in section 3.

1.1 The developing CISP metadata

Our approach is to exploit an ontology methodology with its coherent logic, clear semantics and explicit definitions of the elements used for the development of metadata to describe scientific papers.

Our analysis of papers is based on an ontological representation of investigations. A scientific paper is one of many ways (the most typical) of representing investigations. Our main assumption is that a scientific paper as a representation of the content of a scientific investigation needs to contain the key concepts for the description of investigations. First, we identified what concepts are essential for the description of scientific investigations. Second, we proposed a set of the most essential concepts for representing scientific papers – the Core Information about Scientific Papers (CISP). The principle difference between other metadata schemas and CISP is that the latter aims to represent not just what is typically reported in scientific papers, but what *should be* reported to convey a complete scientific investigation.

We restrict our set of investigations to ones where the research is driven by experimental methods. Our understanding of an experimental method is broad: physically executing experiments, computationally running experiments, or theoretical experiments. Section 4 has a detailed description of CISP classes.

1.2 An ontology engineering

Ontology engineering is still a relatively new research field. Therefore, many of the steps in designing an ontology remain unformalized and can be considered an “art” [Schulze-Kremer 2001]. We give a brief

[1] http://www.ukoln.ac.uk/metadata/

[2] http://uk.dublincore.org/

[3] http://www.textuality.com/mcf/NOTE-MCF-XML.html

[4] http://www.w3.org/PICS/

[5] http://www.techxtra.ac.uk/rss_primer/

description of the basic ideas of an ontology methodology in section 2. We explain the major components of typical ontologies as well as the principles and different approaches to ontology design.

4. CISP classes and underlying ontology

4.1 CISP notation

In this section we follow a format familiar to the reader for describing metadata terms. However many of the listed properties are already built into OWL. CISP follows the naming convention proposed by MSI (The Metabolomics Standards Initiative) working group: <class of instances> [6].

Each CISP class has the following mandatory properties:

- Definition (and Definition reference);
- Location in paper (value = # paper section);
- Representation (default value = natural language text).

Each CISP class also has the following properties and we list them in the description of the CISP classes, even though they are already built into OWL:

- Class name;
- Class ID;
- Cardinality;
- Parent class;
- Subclass;
- Comment.

CISP class can also have the following elements:

- Informal explanation;
- Example;
- and other properties.

CISP includes eight key classes. CISP applications assume mark-up of all these classes in the papers. Many of the key classes have subclasses and properties. They are given in CISP for the better understanding of the semantic meaning of instances that can be found in papers. Mark-up of papers that uses the subclasses and properties of the key classes would provide more semantics to the annotation, but it is not required.

[6] http://msi-ontology.sourceforge.net/

For example, if a research goal in the paper is marked-up as <discover-goal>, not just <goal of investigation>, it would provide more semantics about the nature of the goal and about the investigation. Because <discover-goal> is a subclass of the class <goal of investigation>, any computer system will be able to map this mark-up to the key CISP classes.

4.2 The key CISP classes

We propose the following eight key classes for the description of scientific papers:

- Goal of investigation;
- Motivation;
- Object of investigation;
- Research method;
- Experiment;
- Observation;
- Result;
- Conclusion.

Each class is described in detail in the following sections.

4.2.1 Goal of investigation

Class name:	<goal of investigation>
Class ID:	CISP 1
Cardinality:	multiple
Informal Explanation:	What does an investigation aim to show? What problem does it aim to solve?
Definition:	a goal of an investigation is the target state of the investigation where intended discoveries are made, approaches are tested, problems are demonstrated, tasks formulated etc.
Definition reference:	CISP
Comment:	the definition is derived from the definition of the parent class <goal>

Parent class:	<goal>
Definition:	“A goal is the state that a plan is intended to achieve and that (when achieved) terminates behaviour intended to achieve it”.
Definition reference:	WordNet: http://www.cogsci.princeton.edu/cgi-bin/webwn2.0?stage=1&word=goal+
Subclass:	<confirm-goal>, <explain-goal>, <demonstrate-goal>, <discover-goal>, <observe-goal>, <compute goal>
Example:	development a new approach
Location in paper:	paper section: introduction
Representation:	natural language text FOL (first order logic)

In the literature the most typical goals of an investigation are: to ascertain, to establish, to venture, to discover, to investigate, to infer a fact or theories, the existence of something; to examine, to verify, to falsify, to test a hypotheses, theories, ideas, causal relationships; to gather, to take measurements, to observe data, facts, and other outcomes, etc.

We summarize the goals of an investigation in the list:

1) to check or support a theory on an axiom-deductive basis (a theory-driven approach);
2) to discover a cause-effect dependency on an abductive/ inductive basis (it is not driven by a theory, but a theory might be suggested to explain the experiment results);
3) to demonstrate a known truth (Aristotelian investigation [Medawar, P.B.]);
4) to “find out what happens” in “artificially created situation which allows researcher to manipulate variables” (it is an experiment in Baconian understanding [Medawar, P.B.]);
5) to observe a phenomena;
6) to compute values of an entity of interest.

We call these goals corresponingly: *to confirm, to explain, to demonstrate, to discover, to observe, and to compute*.

We define the follows subclasses of the class <goal of investigation>:

- <confirm-goal>
  Definition: a confirm-goal is a goal of an investigation that uses hypothesis-driven experiments to achieve the goals

of the investigation.

Definition reference: EXPO

- <explain-goal>
  Definition: an explain-goal is a goal of an investigation that uses hypothesis-forming experiments to achieve the goals of the investigation.

  Definition reference: EXPO

- <demonstrate-goal>
  Definition: a demonstrate-goal is a goal of an investigation that uses Aristotelian experiments (to demonstrate a known truth) to provide evidence for already known knowledge.

  Definition reference: CISP

- <discover-goal>
  Definition: a discover-goal is a goal of a Baconian investigation (find out what happens).

  Definition reference: EXPO

- <observe-goal >
  Definition: an observe-goal is a goal of an investigation that uses a descriptive method of research.

  Definition reference: CISP

- <compute goal>
  Definition: a compute-goal is a goal of an investigation that uses computational experiments to achieve the goals of the investigation.

  Definition reference: EXPO

Representation of a goal is usually natural language text, but it can be a logic expression like in the case of the Robot Scientist experiments [King *et. al.,* 2004].

An investigation can have several different goals and a goal usually can be decomposed into sub goals. For example if an investigation includes three experiments, then the goal of the investigation can be decomposed into a set of sub-goals, and each of them constitutes a goal of the experiment. Example: a goal of an investigation is '*development of a new approach'.* In this example, the goal of the investigation is to achieve the target state of the investigation where a new approach has been developed. This goal can be decomposed onto sub-goals: to test the method properties, to compare to other methods, etc.

4.2.2 Motivation

Class name:	<motivation of investigation>

Class ID:	CISP2
Cardinality:	multiple
Informal Explanation:	Why is an investigation important?
Definition:	Motivation for an investigation is the stimulus for achieving the goal of the investigation, the reason to carry out the investigation.
Definition reference:	CISP
Comment:	the definition derived from the definition of the parent class concept <motivation>
Parent class:	<motivation>:
Definition:	The (conscious or unconscious) stimulus for action towards a desired goal, esp. as resulting from psychological or social factors; the factors giving purpose or direction to human or animal behavior. Now also more generally (as a count noun): the reason a person has for acting in a particular way, a motive.
Definition reference:	OED
Subclass:	<application to new domain>
Example:	application of the method to quantum mechanics
Location in paper:	paper section: introduction
Representation:	natural language text

Motivation for the investigation is often stated in the introduction and discussion sections of a paper. There can be several motives.

4.2.3 An object of the investigation

Class name:	<object of investigation>
Class ID:	CISP3
Cardinality:	multiple
Informal Explanation:	About which entities do we seek to gain new knowledge, by means of the investigation?
Definition:	An object of an investigation is the principal entity on which the investigation is based.
Definition reference:	Based on The Oxford English Dictionary. Oxford University Press, 2 Ed., 1989
Parent class:	<role>

	role holder: <method>, <process>, <object>
Example:	model of yeast metabolism (an instance of the class <abstract object>)
Advantage property of object: Informal Explanation:	features of the object that give advantages in certain situations compared to other objects
Disadvantage property of object: Informal Explanation:	features of the object that give disadvantages in certain situations compared to other objects
Location in paper:	paper section: introduction, methods
Representation:	natural language text; link to a classification system of objects

The word ‘object’ in the name of the class should not be understood directly. It does not mean that an object of an investigation is necessarily some physical object. The class <object of investigation> represents what an investigation is about. An object of an investigation is regarded as a role played by some entity (a role holder) in a particular situation: an investigation. Different entities can play a role of an object of an investigation: a method, a process, an animal, a robot (see [Sunagawa *et. al.,* 2005]) for more details about roles). Consider an example where an object of an investigation is a robot. The investigation can be about testing of how this robot is robust and efficient in extreme situations. But in other investigations the same robot can play a role of equipment, e.g. to collect some samples. Consider an example where an object of an investigation is a new method. A class <method> is ontologically defined as <proposition>, not as <object>, but method can have different roles (be a role holder) including <object of investigation>. An investigation about a new method would include studying and testing properties of the method (i.e. accuracy, efficiency), demonstration of areas of application, comparison of its performance to other methods.

An investigation can have several objects of investigation and is often classified according to the object (and it is another role: a base of a classification system).

4.2.4 Method of the investigation

Class name:	<research method>
Class ID:	CISP4

Cardinality:	multiple
Informal Explanation:	How did the authors achieve the goal of the investigation?
Definition:	A research method is a way to solve a scientific task "based upon or regulated by science, as opposed to mere traditional rules or empirical dexterity."
Definition reference:	The Oxford English Dictionary http://dictionary.oed.com
Parent class:	<method>
Subclass:	<experimental method>, <analytical method>
Advantage of method: Informal Explanation:	features of the method that give advantages in certain situations compared to other methods
Disadvantage of method: Informal Explanation:	features of the methods that give disadvantages in certain situations compare to other methods
Example:	the DNA detection method
Location in paper:	paper section: methods; methods and materials
Representation:	natural language text; protocol

CISP includes the following research methods:

- <experimental method>
  Definition: An experimental method is a way to solve a scientific task by designing and executing a scientific experiment.

  Definition reference: EXPO

- <analytical method>
  Definition: an analytical or theoretical method is a way to solve a scientific task by operating with abstract entities like models, theories on deductive basis.

  Definition reference: CISP

What is the difference between an experimental method and other scientific methods such as analytical (or theoretical) methods, descriptive methods, and quasi-experimental methods? The value and power of experimental methods derive from the fact that they allow researchers to detect the laws (cause-and-effect relationships) of nature. On the contrary, descriptive research methods describe phenomena as they occur without aiming to manipulate or control the phenomena in order to establish cause-and-effect relationships (Davis, J.). Examples of descriptive methods are naturalistic observation and case study.

4.2.5 Experiment

Class name:	<experiment>
Class ID:	CISP5
Cardinality:	multiple
Informal Explanation:	If the method of investigation is experimental, what types of experiments were executed? What were the experimental conditions, controls, protocols?
Definition:	A scientific experiment is a procedure which permits the investigation of cause-effect relations between known and unknown (target) variables of the domain. Experimental results cannot be known with certainty in advance.
Definition reference:	EXPO
Subclass:	<computational experiment>: <computer simulation>; <physical experiment>: <Baconian experiment>, <Galilean experiment>: <hypothesis-driven experiment>, <hypothesis-forming experiment>
Example:	
Location in paper:	paper section: methods
Representation:	natural language text; protocol

CISP supports the following subclasses of the class <experiment>:

- <computational experiment>
  Definition: A computational experiment is a scientific experiment which investigates cause-effect relations between known and unknown (target) variables by manipulating the computational (non-physical) domain adequate to the real-world domain.

Definition reference: EXPO

- <computer simulation>
  Definition: Computer simulation is a computational experiment where the real domain of study is modelled by a computer program imitating “the internal processes and not merely the results of the thing being simulated".

  Definition reference: The Free Dictionary: http://www.thefreedictionary.com/computer+simulation

  Parent class: <computational experiment>

- <physical experiment>
  Definition: A physical experiment is a scientific experiment which investigates cause-effect relations between known and unknown variables by manipulating the real-world (physical) domain.

  Definition reference: EXPO

- <Baconian experiment>
  Definition: A Baconian experiment is a scientific experiment which involves no explicit hypothesis.

  Definition reference: Medawar, P.B. Advice to a Young Scientist. Pan Books Ltd, London, 1981

  Parent class: <physical experiment>

- <Galilean experiment>
  Definition: A Galilean experiment is a scientific experiment which involves explicit hypotheses.

  Definition reference: Medawar, P.B. Advice to a Young Scientist. Pan Books Ltd, London, 1981

  Parent class: <physical experiment>

- <hypothesis-driven experiment>
  Definition: A hypothesis-driven experiment is a Galilean experiment designed to confirm or reject a given hypothesis. The deductive consequences of the hypothesis are compared with the experimental result, and the probability of the hypothesis is either increased (confirmed) or decreased (rejected).

  Definition reference: EXPO

  Parent class: <Galilean experiment>

- <hypothesis-forming experiment>
  Definition: A hypothesis-forming experiment is a Galilean experiment. It includes a hypotheses formation stage in which one or more hypotheses are formed

using abduction or induction.

Definition reference: EXPO

Parent class: <Galilean experiment>

The class <experiment> can additionally have the following properties:

- *has Protocol*, where the class <protocol> is defined as:
  Definition: An experiment protocol is an explicit detailed specification of an experiment which describes a plan of experiment actions to achieve an experiment goal.

  Definition reference: EXPO

  Parent class: <procedure>

- *has Equipment*, where the class <equipment> is defined as:

  Definition: Experiment equipment is the set of tools, devices, materials, computer systems assembled for performing the experiment.

  Definition reference: based on Collins Softback English Dictionary, HarperCollins Publishers, Glasgow, 1993.

- *has Design*, where the class <design> is defined as:

  Definition: An experiment design is a structured, organized method for determining the relationship between factors affecting cause-effect relations between known and unknown variables.

  Definition reference: Sigma. http://www.isixsigma.com/dictionary/Design_of_Experiments_-_DOE-41.htm

- *ha sExperiment Factor*, where the class <experiment factor> is defined as:

  Definition: An experiment factor is a known variable of the model of the domain which the object of the experiment can control/vary in order to determine of a value of target variables.

  Definition reference: EXPO

- *ha sObservation*, where the class <observation> is a key CISP class and defined in the section below.

- *has Result*, where the class <result> is a key CISP class and defined in the section below.

The formal description of experiments for efficient analysis, annotation, and sharing of results is a fundamental part of the practice of science. The above listed properties are important for providing information about the experiments executed within the investigation reported in a paper: what type of experiments were executed, what were the factors, how they were designed, what equipment was used, characteristics of the latter etc. CISP does not require mark-up of these properties, but it will add more semantics to the paper's mark-up output. There are several ontology-based projects in bio-medical domains investigating how to record such information[7,8]; and metadata standards are appearing in many other sciences, e.g. in Physics[9]. Probably the best known attempt to formalise the description of experiments is that developed by the Microarray Gene Expression Society (MGED)[24]. The MGED Ontology (MO) was designed to formalise the descriptors required by MIAME (Minimum Information About a Microarray Experiment) standard for capturing core information about microarray experiments. MO aims to provide a conceptual structure for microarray experiment descriptions and annotation. A number of ontological developments related to MO also exist. The HUPO PSI General Proteomics Standards and Mass Spectrometry working groups are building an ontology that will support proteomic experiments[25]. The MSI (Metabolomics Standards Initiative) ontology working group is seeking to facilitate the consistent annotation of metabolomics experiments by developing an ontology to help enable the scientific community to understand, interpret and integrate metabolomic experiments[10]. More generally, the Functional Genomics Investigation Ontology (FuGO), now is known as OBI (an ontology of bio-medical investigations) is developing an integrated ontology that provides both a set of "universal" terms, i.e. terms applicable across functional genomics, and domain-specific extensions to terms[11].

CISP aims to provide semantic mark-up for investigations from various domains, while being consistent with already existing (or in progress) representations for specific domains, like OBI.

4.2.6 Observation

Class name:	<observation>
Class ID:	CISP6
Cardinality:	multiple
Informal Explanation:	What data/phenomena were recorded within an investigation? How are the data represented, in what format, where are they stored?
Definition:	An experimental observation is a direct observation of nature, the set of values of target variables (or other variables of the domain), "prior to analysis; interpretation*" (compare with

[7] mged.sourceforge.net/

[8] psidev.sourceforge.net/

[9] www.ph.ed.ac.uk/ukqcd/community/the_grid/QCDml1.1/ConfigDoc/ConfigDoc.html

[10] msi-ontology.sourceforge.net/index.htm

[11] fugo.sourceforge.net/

	results).
Definition reference:	Based on The Oxford English Dictionary. Oxford University Press, 2 Ed., 1989 * Collins Softback English Dictionary, HarperCollins Publishers, Glasgow, 1993.
Example:	Optical density readings
Location in paper:	paper section: results and discussion
Representation:	natural language text
External location:	data base

The class <observation> can provide an explicit link between the paper and the data. The semantic mark-up of the paper in future aims to contain all necessary information about the data stored in data base or other public resource: location, access rights, version, supporting systems, etc.

4.2.7 The results

Class name:	<result>
Class ID:	CISP7
Cardinality:	multiple
Informal Explanation:	What are the main outcomes of an investigation?
Definition:	results of the investigation are the set of facts, obtained through the interpretation of the observations.
Definition reference:	EXPO
Subclass:	<experiment result>
Example:	An average curve representing the growth of wild type yeast
Location in paper:	paper section: results and discussion
Representation:	natural language text
External location:	data base

The class <Results> (as does <Observation>) can provide an explicit link between the paper and the data. The semantic mark-up of the paper in future will contain all necessary information about the results stored in the data base or other public resource: location, access rights, version, supporting systems, etc.

The distinction between the classes <Result> and <Observation> is debatable. Many researchers consider them as synonyms. We would like to stress within the CISP formalism the difference between direct observations of some phenomena and processed or interpreted data. The status of such data is different from the point of view of the level of evidence they both provide towards reaching conclusions.

4.2.8 Conclusion

Class name:	<conclusion>
Class ID:	CISP8
Cardinality:	multiple
Informal Explanation:	What new knowledge has been discovered? Has the goal of an investigation been achieved? Has a hypothesis been confirmed?
Definition:	A conclusion of an investigation is a statement inferred from observations, results, assumptions, and facts to support or reject a research hypothesis.
Definition reference:	EXPO
Example:	A particular gene has a particular function
Location in paper:	paper section: abstract, conclusion, discussion
Representation:	natural language text

It is important to record under what assumptions, restrictions the conclusions were made; what facts and evidences are there to support such conclusions.

4.3 Candidate classes for inclusion into CISP

The following classes were considered for inclusion into the list of the key CISP classes:

- <hypothesis>

  Informal Explanation: What is a hypothesis of an investigation, an experiment, or model? What was tested in the investigation?

  Definition: A hypothesis is a statement about cause-effect relations between known and unknown (target) variables of the domain of the investigation "that shall be in accordance with known facts" to be verified by the experiment.

Definition reference: The Oxford English Dictionary. Oxford University Press, 2 Ed., 1989

Example: A particular yeast strain will have a higher growth rate than a wild strain.

- <background fact>

Informal Explanation: A neutral or widely accepted statement about the knowledge domain.

Example: The addition of benzotriazole dyes to oligonucleotides either using a dye phosphoramidite, or using post oligonucleotide synthesis via suitable linker has been reported [#ref]. The post synthetic methods are easier to use in practice although lower yields are obtained.

<problem>

Informal Explanation: The difficulties, restrictions when trying to achieve the goal of an investigation.

Example: DNA does not meet the requirements for SERRS due to the lack of a suitable visible chromophore.

- <example>

Informal Explanation: The examples given to demonstrate the authors' findings or to explain an approach.

- <model>

Informal Explanation: What theoretical model was used within an investigation? How is the model represented: as a system of equations, as logical rules and facts?

Example: Logical model of yeast metabolism

- <domain of investigation>

Informal Explanation: To what area of research an investigation belongs. To what area of knowledge the main results contribute.

Example: Physical Chemistry

Map to: <DC: Subject>

We would like to give some remarks about the class <domain of investigation>. This class is not directly used for the annotation of papers with CISP, and DC has a similar term <DC: Subject>. CISP application needs to ‘know’ what the domain of the annotated paper is, in order to download corresponding domain ontologies and data bases. It is not a trivial task to define what a domain of an investigation is. There can be several domains that are associated with an investigation. We suggest dividing them onto two main classes:

- <main domain of investigation>

  Informal Explanation: They are of research in which new knowledge is discovered

  Parent class: <domain>

- <supplementary domain of investigation>

  Informal Explanation: associated/ auxiliary domain

  Parent class: <domain>

The following classification system will be additionally used by our system:

- <research councils UK classification>

  Informal Explanation: The research councils UK classification is a classification of domains used by Britain's councils for academic research.

4.4 Verification of CISP

In order to define CISP, the Core Information about Scientific papers, we first chose a subset of the general scientific concepts (GSCs) described in EXPO. Our choice was based on the results of interviews with the experts and preliminary annotation of papers. Three experts were asked to annotate four papers using this set and at the end of this preliminary annotation the list of concepts was refined to include the following: **Goal of Investigation, Object of Investigation, Method of Investigation, Experiment, Observation, Hypothesis, Results, Conclusion, Motivation, Background, Problem, Example**.

The reasoning behind this choice was to determine a set of concepts that would describe the scientific investigation represented in the paper as integrally as possible, in terms of its objectives, the methodology of the approach, the outcomes and the pre-existing work which sets the scene for the current investigation.

Before proceeding on a large scale annotation of papers using the above concepts we wanted to assess whether the research community felt in agreement with our constructed set of terms and which of the GSCs were considered most informative and therefore indispensable.

We conducted an online survey (not anonymous) where each concept was presented as a candidate for inclusion in CISP, along with a short definition and an example of its use. The survey is available at:

http://www.aber.ac.uk/compsci/Research/bio/art/news/survey/

and everyone is invited to answer on our questions about CISP. We asked participants from a range of research disciplines from the UK and Japan to vote for the concepts they thought should be part of CISP. At the time of writing we have received feedback from 33 researchers.

We have ranked below the GSCs in descending order of popularity according to the survey participants. The five highest scoring concepts and the short description given on the website are:

1) Conclusion (the new knowledge that has been discovered or whether the goals of an investigation have been met);
2) Results (the actual outcomes of the investigation);
3) Goal of the investigation (what the investigation aims to show);
4) Method of the investigation (how the authors set out to pursue the goals of the investigation);
5) Object of the investigation (entity about which we seek to gain new knowledge).

The above are followed by <Experiment> (the types of experiments executed) and <Observation> (what data/phenomena were recorded within an investigation).

These were mostly the concepts from the proposed CISP. However, judging from the votes obtained, <Experiment> and <Observation> are not considered as significant in describing the methodological approach and outcomes of the investigation respectively.

In addition, Motivation which we considered to be highly relevant to an investigation, did not score as many votes. For example, some researchers saw it as being hard to distinguish from <Goal of an investigation>. Two thirds of our participants thought we should also include the hypothesis of an investigation in CISP but in general they hold concepts pertaining to pre-existing work and the state of the art <Problem>, <Background> to be less crucial. Finally, the theoretical <Model> employed and any examples used to illustrate the approach are considered by most to be too detailed aspects of the methodology.

A comment made by several of our participants was that our analysis may not be suited to all kinds of scientific papers, for instance review papers or papers that simply showcase systems may lack most of the concepts in CISP. We are taking this on board and are planning to focus on papers that are original contributions with a determinable investigation and results. Also, many people added that not all science has an explicit hypothesis either. However, even if the hypothesis of an investigation is not stated as such in a scientific paper, it can almost always be inferred, which is what we plan to do in our analysis. Another participant mentioned how they often find there is an overlap between <Goal>, <Object> and <Method>. It is often the case that one concept may subsume the other, e.g. sometimes developing a method can also be the object of an investigation but the details of the method are interesting in their own right. Nevertheless, we will make sure the distinction between the latter three concepts is as clear as possible to our annotators.

Overall we have found this survey to be very useful as it has helped us obtain a third party view on CISP and the importance of GSCs. Many of the comments have been invaluable and we will definitely take them into account. The survey results are summarized in the table (see the Appendix).

5. Conclusion

We have proposed the CISP metadata for the description of papers about scientific investigations. The principle difference between CISP and other metadata schemas is that the former aims to represent not just what is typically reported in scientific papers, but what *should be* reported to convey a complete scientific investigation. We restricted our set of investigations to ones where the research is driven by experimental methods. Our understanding of an experimental method is broad: physically executing experiments, computationally running experiments, or theoretical experiments.

CISP is a defined set of leaf classes from DOLCE and EXPO. CISP has eight key classes: <Goal of investigation>, <Motivation>, <Object of investigation>, <Research method>, <Experiment >, <Result>, <Observation>, and <Conclusion> and we have provided the detailed description of each CISP class.

CISP has sound theoretical foundations. The methodology used to construct CISP, ensures logic coherence and compliance to other formalisms. As so far, CISP has been validated by experts in physical chemistry and by researchers from different fields who participated in an online survey. Within the ART project we plan to apply CISP to a set of 200 papers to verify our approach.

We argue that ontologies are important for the development of metadata. Ontology engineering provides semantic clarity, explicitness, and facilitates the reusability of represented information and knowledge. The use of formalized semantic representation can also facilitate natural language processing for intelligent information analysis and retrieval. Therefore ontology based knowledge representation opens new perspectives for text mining techniques and logic inference.

Ontologies with their explicit definitions and clear structure are also a valuable resource for educational applications.

About the authors

Dr Larisa Soldatova is a RCUK Fellow at the University of Wales, Aberystwyth, Department of Computer Science. Over the past ten years, she has worked on ontology development and knowledge representation in Russia, Japan and the UK. The focus of her research is the development of formal methods for knowledge representation and their applications to the natural and social sciences. She developed EXPO, a general ontology of scientific experiments. So far she has applied EXPO to biology, physics, computer science, data base design, and has developed an ontology of tests (exams).

Larisa has a PhD in computer science from the Far Eastern Technical University (Russia) on the topic of automated generation of problems.

She can be contacted at: lss@aber.ac.uk.

Dr Maria Liakata has been a research associate with the Computational Biology group at the University of Wales, Aberystwyth since 2005. She is a computer scientist with a mathematical and Natural Language Processing (NLP) background. Maria's research interests include Computational Semantics, machine learning for knowledge discovery and the use of ontologies in NLP applications. She has also worked on data mining from /*Arabidopsis*/ mass spectrometry data and on the computational modelling of yeast growth curves.

Maria has a DPhil in Computational Linguistics from the University of Oxford on the topic of Inducing Domain Theories.

She can be contacted at: mal@aber.ac.uk.

References

Boniface, D. R. (1995) Experiment design and statistical methods for behavioural and social research. London: Chapman & Hall.

Curd, M., & Cover, J. A. (1988) Philosophy of Science. W.W. Norton & Company.

Fielding J.M., Simon J., Ceusters, W., Smith, B., 2004, Ontological Theory for Ontological Engineering: Biomedical Systems Information Integration. In Proc. The principles of knowledge Representation and Reasoning.

Fisher, R. A. (1956) The design of experiments. Oliver & Boyd, Edinburgh.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L. Sweetening Ontologies with DOLCE. In A. Gómez-Pérez, V.R. Benjamins (eds.) Knowledge Engineering and Knowledge Management. Ontologies anHobbs, J.R., Feng, P., (2004) An Ontology of Time for the Semantic Web. ACM Transactions on Asian Language Processing (TALIP): Sd the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Springer Verlag, pp. 166-181.

Hobbs, J.R., Feng, P., (2004) An Ontology of Time for the Semantic Web. ACM Transactions on Asian Language Processing (TALIP): Special issue on Temporal Information Processing, 3/1: 66-85.

King, R. D., Whelan, K.E., Jones, M.F., Reiser, P.G.K, Bryant, C.H. (2004) Functional Genomics Hypothesis Generation by a Robot Scientist. Nature, 427, no 6971, 247-252.

Kozaki, K., Kitamura, Y., Ikeda, M. & Mizoguchi, R., 2002, Hozo: An Environment for Building/Using Ontologies Based on a Fundamental Consideration of "Role" and "Relationship". Knowledge Engineering and Knowledge Management, 213-218.

Medawar, P.B. (1981) Advice to a Young Scientist. Pan Books Ltd, London.

Mizoguchi, R., 2004a, Tutorial on ontological engineering - Part 2: Ontology development, tools and languages. New Generation Computing, OhmSha&Springer, 22/1: 61-96.

Mizoguchi, R., 2004b, Tutorial on ontological engineering - Part 3: Advanced course of ontological engineering. New Generation Computing, OhmSha&Springer, 22/2: 193-220.

NISO (2004) Understanding Metadata. Bethesda, NISO Press.

Rosse, C., Mejino Jr., 2003, A reference ontology for biological informatics: the Foundational Model of Anatomy. Biomedical Informatics, 36, 478-500.

Schierz, A.C., Soldatova, L. N. and King, R.D. (2007) Reply to Overhauling the PDB. Nature Biotechnology 25/8: 846.

Schulze-Kremer, S., 1997, Proc. Int. Conf. Intell. Syst. Mol. Biol. 5, 272-275

Schulze-Kremer, S., 2001, Ontologies for Molecular Biology. Computer and Information Sci. 6(21)

Smith, B., 2003, The Logic of Biological Classification and the Foundations of Biomedical Ontology. Dag Westerståhl (ed.), (Invited paper). In: Proc. 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain.

Soldatova, L.N., King, R.D., 2005, Are the Current Ontologies used in Biology Good Ontologies? Nature Biotechnology 9/23 pp. 1096-1098

Soldatova, LN & King, RD. (2006) An Ontology of Scientific Experiments. Journal of the Royal Society Interface 3/11: 795-803.

Sowa, J.F. (2000) Knowledge Representation. Logical, Philosophical, and Computational Foundations. Brooks/Cole.

Sunagawa, E., Kozaki, K., Kitamura, Y., Mizoguchi, R., 2005, A Framework for Organizing Role Concepts in Ontology Development Tool: Hozo. AAAI Symposium Roles, an Interdisciplinary Perspective: Ontologies, Programming Languages, and Multiagent Systems, USA, FS-05-08, pp.136-143.

The Collins Softback English Dictionary, HarperCollins Publishers, Glasgow, 1993.

The Oxford English Dictionary, 1989, Oxford University Press, 2 Ed.

Toulmin, S. (2004) Philosophy of Science. in Encyclopaedia Britannica, Deluxe CD.