ART: An ontology based tool for the translation of papers into Semantic Web format

Larisa N. Soldatova[1], Colin R. Batchelor[2], Maria Liakata[1], Helen H. Fielding[3], Stuart Lewis[1] and Ross D. King[1]*

1) University of Wales, Aberystwyth, UK; 2) Royal Society of Chemistry Publishing, UK; 3) University College London, UK

ABSTRACT

The paper describes initial work on an ontology based tool, ART, for the semantic annotation of papers stored in digital repositories. ART is intended for the annotation not only of data and metadata about a paper, but also the main elements of the described scientific investigation, such as goals, hypotheses, and observations. ART will also be able to aid in the expression of research results directly in a semantic format, through the composition of text using ontology-based templates and stored typical key phrases for the description of basic elements of the research. ART's system design, its functionality, and related projects are discussed. An example annotated paper is presented in order to demonstrate the expected output of the tool.

1 INTRODUCTION

Semantic Web technologies use semantic metadata to improve information retrieval and knowledge representation. Metadata provide semantic clarity, explicitness, and facilitate the reusability of represented information and knowledge [Soldatova & King, 2006a]. Ontology based semantic annotation of papers and data promotes the sharing of research results, and reduces the duplication and loss of knowledge. It also facilitates text mining and knowledge discovery applications.

We are developing ART (an ontology based ARticle preparation Tool), a practical annotation tool which can be used to add value to repository papers and data. ART will generate annotations containing not only metadata about the paper (title, author, etc.), but also generic scientific concepts, such as the type of investigation (theoretical or experimental), its goal, results, the reliability of the results, etc. ART also aims to automate the recognition of those generic concepts in a text. The tool will use a number of Open Biomedical Ontologies (OBO) (http://obi.sourceforge.net/) to find in the text domain specific concepts and to link them to external sources. The result will be an article in OWL-DL (http://www.w3.org/TR/owl-guide/) format that can be submitted to a digital repository along with the original article free-text. The OWL version of the article could then be used for a variety of computational applications (e.g. data mining); or by researchers to check explicit explanations of some terms from the text, or to get more details about experiments.

We also envisage ART being used by authors at the time of manuscript submission to generate annotations, expressed in OWL, that describe the paper and related data. The tool will lead the author through a process where: experimental goals, hypotheses, methodologies, and results, are described and linked to the text and external data files. ART will check that all required information is present, and if necessary give examples of formulating concepts.

As a part of the project we will create a digital repository of papers in OWL format. This repository will be an example of an intelligent digital repository. It will be possible to use it for investigation of advanced text mining and knowledge discovery techniques. Since all the papers will be represented in enriched semantic format and directly linked to data sources, new intelligent queries, like: “find evidence for the given hypothesis”, “is the research conclusion consistent with the evidence and the assumptions?” will be possible.

The rest of the paper is organised as follows: section 2 has a brief description of the related work; section 3 presents the ART project, its goals, tasks and principles; section 4 describes the design of the ART tool, main modes and functions; an example of an annotated paper is considered in section 5; whereas section 6 discusses the current state of the project and future plans.

2 RELATED PROJECTS AND LINKS

The ART project aims to build on the experience of eBank. The later project aims to provide a technological solution to the access and curation of digital resources (http://www.ukoln.ac.uk/projects/ebank-uk/). The project is being led by UKOLN in partnership with the Intelligence, Agents & Multimedia Group, Department of Electronics & Computer Science, and the Department of Chemistry, University of Southampton and the Digital Curation Centre (DCC) (http://www.dcc.ac.uk/). eBank/e-Print already uses Dublin Core Metadata (DC) (http://dublincore.org/) to index the e-prints and enable searching. The ART project will

contribute to the DCC services. The DCC priorities related to ART are:

- Metadata extraction and curation (investigating standards and tools for the curation of scientific metadata).
- Semantic data curation ('meaning' and 'machine process-ability' foundations of the Semantic Web and Ontological communities).
- Data transformation, integration and publishing (manipulation of data formats, metadata conversion).

The ART project will help to ensure that metadata for various scientific domains are stored and updated in one place. It can provide consistency in managing the digital resources.

ART will build on the experience of semantic enrichment with the RSC's (Royal Society of Chemistry) Project Prospect (http://www.projectprospect.org/). Here journal articles are marked up with chemical structures and domain terms from the IUPAC Gold Book [International Union of Pure and Applied Chemistry, 1997] and terms from the OBO ontologies GO (Gene Ontology) [The Gene Ontology Consortium, 2000], SO (Sequence Ontology) [Eilbeck *et al.*, 2005], and CL (Cell Type ontology) [Bard *et al.*, 2005]. Mark up of terms from ChEBI (Chemical Entities of Biological Interest) [Matos *et al.*, 2006], FIX (ontology of physico-chemical methods and properties) (http://obo.sourceforge.net/cgi-bin/detail.cgi?fix) and REX (ontology of physico-chemical processes) (http://obo.sourceforge.net/cgi-bin/detail.cgi?rex) is in preparation. ART will also incorporate generic scientific concepts from EXPO (Ontology of scientific EXPeriments) [Soldatova & King, 2006b], OBI (Ontology for Biomedical Investigations) (http://obi.sourceforge.net/), ECO (Evidence Code Ontology) (http://obo.sourceforge.net/cgi-bin/detail.cgi?evidence_code).

A similar ontology based format is going to be used for the related ROAD (Robot-generated Open Access Data) project (http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/road.aspx). This project will be investigating the issues involved with the automatic routine deposit of data generated by the Robot Scientist [King *et al.*, 2004].

ART aims to advance digital repositories technology, and provide an ontological foundation for manuscript annotation and formalisation.

3 ART PROJECT

The aim of the JISC (Joint Information Systems Committee, UK) funded ART project is to develop an ontology based tool to assist in:

- Translating scientific papers into a format with an explicit semantics.
- Explicit linking of repository papers to data and metadata.
- Creation of an example intelligent digital repository.

We would like to stress that the generic approach and further extensibility of the tool are core principles for the system design. The article translation and preparation tool ART is intended to eventually be a general purpose for any domain where there are available ontologies. The domain independent parts of the system will be fully reusable, and the domain dependent part must be provided with the list of external domain sources. The restriction is that for some domains formalized representations do not exist yet. However ontology development is a rapidly progressing area.

To develop ART we are currently focusing on physical chemistry as the application domain. The rationale for this is that chemistry publications are among the most formalized of all the sciences, and the eBank project has already used chemistry as an exemplar. In addition, physical chemistry papers employ many concepts that have already been formalized in a number of OBO ontologies, i.e. <CHEBI: molecular entity>, <OBI: solid state>, <SBO: concentration>.

4 DESCRIPTION OF THE SYSTEM

The ART system will have two main parts: domain independent and domain dependent (see fig. 1). Each of these will use corresponding ontologies to annotate text. Domain independent sources such as DC Metadata, EXPO, OBI and ECO will be incorporated into the system. The tool will import domain dependent sources after identification of the domain. For physical chemistry these sources are: ChEBI, FIX, REX, and IUPAC. The system can be easily extended by including more internal and/or external ontologies and other sources.

ART will use natural language processing techniques to support both domain-specific and domain-independent mark up. ART itself will use SciXML [Rupp et al., 2006] to represent scientific articles. There exists a framework [Hollingsworth et al., 2005] for converting PDF, which is the most likely format for article submissions, into SciXML, which will provide bibliographic metadata such as titles (<DC: title>), authors (<DC: creator>). The domain-specific mark up of OBO concepts can be partly achieved through named entity recognition. [Batchelor and Corbett, 2007]. The automatic identification of domain-independent concepts is significantly more challenging. However, they frequently occur in well-defined 'zones' within articles [Teufel et al., 1999], and are often introduced by meta-discourse markers or 'cue phrases' articles [Teufel, 1998]. The system will use this information to attempt to identify generic concepts such as <EXPO: goal> or <OBI: conclusion> and will ask users to confirm or to correct the identified concepts in

interactive mode. The system will be able to provide a user with explanations why these concepts are necessary, and give definitions and examples. The outcome will be a semantically enriched paper in a text format and OWL paper annotation. If the user wishes, the tool can automatically generate a summary of the article and RSS feed.

ART will also be able to help the author represent his/her research results directly in OWL format. The system will ask for input of the required metadata and data about the research, and will provide examples and explanations where necessary. ART will be designed to assist in composing a paper reporting the results of the investigation. The system will have ontology-based templates of papers. After collecting all the metadata and data about the investigation, the system will propose a paper structure and give examples of key phrases for the description of the main research components.

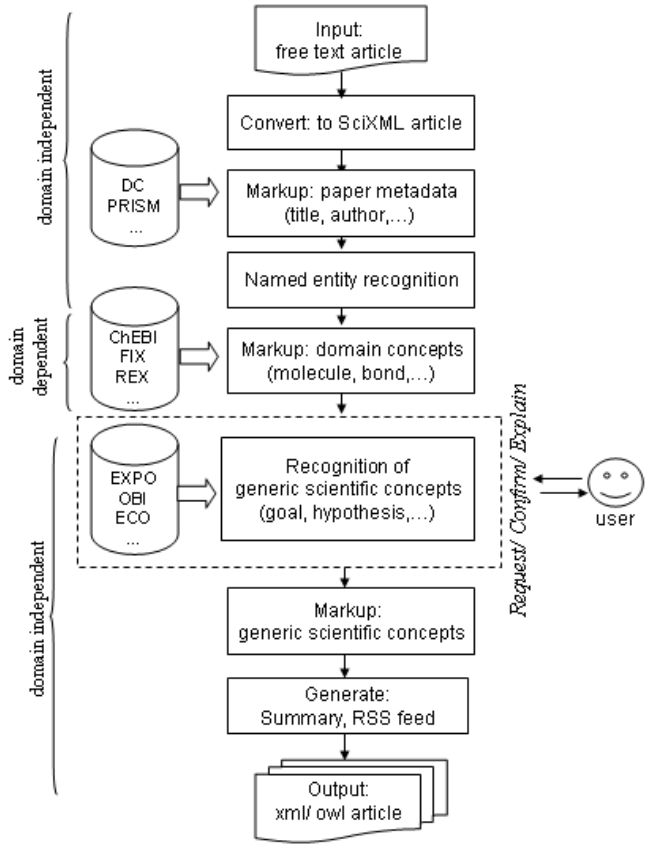

Fig. 1. A flow diagram for the ART system for the physical chemistry domain.

5 EXAMPLE

Let us consider an example physical chemistry paper from the Faraday Discussions [Hiberty *et al.*, 2006]. This paper discusses the notion of a charge shift (CS) bond, and is particularly interesting for our purposes because it is a theoretical computational investigation. This area is at present poorly covered by existing ontologies and we will have to extend EXPO to handle these investigations, which are common in many areas of science. Figure 2 shows a fragment of the annotated scientific concepts in the example paper in order to demonstrate the outcome of the ART tool:

<EXPO: investigation> *map to*
<DC: title> *map to*
<OBI: investigation>
The physical origin of large covalent-ionic resonance energies in some two-electron bonds.

<EXPO: goal>
"studying in detail all the aspects of bond formation in a series of molecules that each display a range of bonding features: H_2 and C_2H_6 as members of the classical family of covalent bonds, Cl_2 as a bond exhibiting significant CS character, and the series N_2H_4, H_2O_2, and F_2 as molecules exhibiting increasing CS character from left to right of the periodic table."

<EXPO: object of investigation> *map to*
<OBI: investigation object role>
characteristics of CS bond in H_2, C_2H_6, Cl_2, N_2H_4, H_2O_2, F_2 molecules

<EXPO: method>
valence bond calculation on two levels:
valence-bond-self consistent field (VBSCF)
breathing-orbital valence bond (BOVB)

<EXPO: method assumption> *map to*
<ECO: traceable author statement>
"all the orbitals, including the inactive set, are kept strictly localized, and the ionic components are described as simple closed-shell VB functions" for H_2, C_2H_6, N_2H_4, H_2O_2, molecules".

<EXPO: experimental equipment>
<OBI: software>
XMVB Program
Gaussian 98 series Program

<EXPO: experiment results>
<EXPO: computational data>
Dissociation energy curves of the purely covalent VB structure for H_2, C_2H_6, Cl_2, N_2H_4, H_2O_2, F_2 molecules
(calculated with VBSCF method)
VB-3 three structure ground state for H_2, C_2H_6, Cl_2, N_2H_4, H_2O_2, F_2 molecules
(calculated with BOVB method)

<EXPO: conclusion> *map to*
<OBI: conclusion>
"CS bonding is characterised by the following features: (i) a covalent dissociation curve with a shallow minimum situated at long interatomic distance, or even a fully repulsive covalent bond; (ii) a large covalent-ionic resonance energy RE_{cs} that is responsible for the major part, or even for totality, of the bonding energy."

Fig. 2. A fragment of the annotated article in a text format.

The system will provide mappings between the incorporated representations. The same element in the text can be linked to a number of internal and/or external resources. For instance in the example considered, the title of the paper is marked as <EXPO: investigation> which corresponds to the class <OBI: investigation> and to the term <DC: title>. These mappings are not equivalent, but ART will contain formalized description of the semantics involved.

Some generic scientific concepts can be automatically recognized by the system using the cue phrases. For instance the phrase in the considered paper "this paper is aimed at…" indicates the goal of the investigation and the metadata about the text structure <section: conclusion> points out to the list of the conclusions. The system will be able to identify more cue phrases for the recognition of more elaborate concepts. It will not always be possible to automatically identify concepts in free text. However the system 'knows' what should be in a scientific paper and can ask a user, for example "What are the results of the investigation?" The user can then indicate them in the text or input them directly.

We expect that the ART tool can also help with ontology construction. For example in the paper considered above, the new concept <CS bond> is discussed, and features of such bonds are investigated. This concept is absent from all existing ontologies. The system could collect such missing terms and then ontology developers would consider them for inclusion to the corresponding ontology.

6 DISCUSSION

The potential users of the ART tool are: curators of digital repositories who would like to semantically enhance the papers stored in the repositories; researchers who would like to represent their research results in semantic machine readable format for various computer applications; publishers and reviewers. Reviewing is a time-consuming process and any tool to facilitate the process will be of significant value.

ART is an ongoing project and the ART tool is in a development stage. The authors would like to use this opportunity and invite potential users for feedback on the proposed functionality of the system.

ACKNOWLEDGEMENTS

The work was funded by JISC (Joint Information Systems Committee, UK).

REFERENCES

Batchelor, C.R. and Corbett, P.T. (2007) Semantic enrichment of journal articles using chemical NER. *In proc. of ACL* (in press).

Hiberty, Ph.C., Ramozzi, R., Song, L., *et al*. (2006) The physical origin of large covalent-ionic resonance energies in some two-electron bonds. *Faraday Discuss*., **135**, 261-272.

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*, **427/6971**, 247-252.

Soldatova, L.N. and King R.D. (2006) Ontology Engineering for Biological Applications. *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences.* Christopher J.O. Baker and Kei-Hoi Cheung (Eds). Springer, NY., 121-137.

Soldatova, L.N. and King, R.D. (2006) An Ontology of Scientific Experiments. *Journal of the Royal Society Interface*, **3/11**, 795-803.

Teufel, S., Carletta, J., Moens, M. (1999) An annotation scheme for discourse-level argumentation in research articles. *In Proc. of EACL*.

Teufel, S. (1998) Meta-discourse markers and problem-structuring in scientific articles, *Workshop on Discourse Structure and Discourse Markers, ACL*, Montreal.

International Union of Pure and Applied Chemistry (1997) Compendium of Chemical Terminology. 2nd edition. Blackwell Science, Oxford.

The Gene Ontology Consortium (2000) Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, **25**, 25-29.

Eilbeck K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M. (2005) The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* **6**, R44.

Bard, J., Rhee, S.Y. and Ashburner, M. (2005) An ontology for cell types. *Genome Biology* **6**, R21.

Matos, P., Ennis, M., Darsow, M., Guedj, M., Degtyarenko K. and Apweiler R. (2006) ChEBI - Chemical Entities of Biological Interest. *Nucleic Acids Research*, Database Summary paper 646.

Rupp, C.J., Copestake, A., Teufel, S. and Waldron, B. (2006) Flexible Interfaces in the Application of Language Technology to an eScience Corpus. *In Proc. of the 4th UK E-Science All Hands Meeting*. Nottingham, UK.

Hollingsworth W., Lewin I. and Tidhar D. (2005) Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a

Prerequisite for E-Science Text Mining. *In Proc. of the 4th UK E-Science All Hands Meeting*, Nottingham, UK, 267-273.