Introducing a content integration process for a federation of agricultural institutional repositories

Vassilios Protonotarios[1,4], Laura Gavrilut[1,4], Ioannis N. Athanasiadis[1,2], Ilias Hatzakis[1], Miguel-Angel Sicilia[3]

[1] Greek Research & Technology Network (GRNET), Greece
[2] Democritus University of Thrace, Greece
[3] University of Alcala, Department of Computer Science, Spain
[4] Agro-Know Technologies, Greece
vprot@aua.gr, lgavrilut@yahoo.com, ioannis@athanasiadis.info, hatzakis@grnet.gr, msicilia@uah.es

Abstract. Aggregating metadata from various sources often raises practical issues, such as incompatibility between the different metadata application profiles (AP) used as well as quality aspects of the metadata used. In the case of repositories hosting agricultural-related content, the existence of various metadata AP with significant differences between them makes the effort of interconnecting these repositories a difficult task. This paper proposes a process for integrating different agricultural scientific content repositories and a workflow that should be followed in the context of populating the repository. Although the proposed solution refers to repositories with agricultural content, the same process can be followed in the case of repositories with different content. This process was proposed under the VOA3R project which is funded by the European Commission's ICT PSP Programme.

Keywords: content integration, aggregation, application profile, metadata

1 Introduction

Despite the existence of a significant number of digital repositories hosting content related to agriculture, any effort towards interconnecting these repositories through metadata aggregators exhibits a high degree of complexity. This is due to the fact that the available resources have been annotated with metadata using application profiles with significant differences (e.g. Dublin Core, IEEE LOM, FAO AGRIS, custom etc.) therefore high incompatibility.

In the case of the VOA3R project (http://www.voa3r.eu), a number of repositories for agriculture and aquaculture will be aggregated and access to their resources will be provided through the VOA3R platform. However, the metadata used for the annotation of the resources in these repositories come from a wide variety of metadata

AP, so there is a need for developing a technical solution for the harmonization of these differences and the usage of a common metadata AP, which will be used in all cases.

As a result, the need for a process which will set up aggregators of content repositories, so that metadata is harmonized and quality-certified is raised. In order to facilitate the aggregation of resources coming from various sources, this paper proposes a process/workflow for the integration of institutional repositories, as well as a scenario illustrating its application.

2 Background

2.1 About VOA3R

VOA3R (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment) is a 36-months CIP-ICT-PSP EU project which aims to improve access to European agriculture and aquaculture research results by using an innovative approach to sharing open access research results. One of the goals of the VOA3R project is to create a federation of scholarly content, by organizing and coordinating the population of the VOA3R repository with content coming from the participating repositories and communities. This content will be described with semantically rich and interoperable metadata [1]. In order to achieve this, a workflow, which should be followed by all the members of the VOA3R federation during the phase of content population, in the VOA3R repository is developed. The scope of the process is to integrate new resources to the existing collections and populate the VOA3R federation with educational agricultural content.

In order to come up with a common and compatible framework for the annotation of resources aggregated from various repositories, thus minimizing any incompatibility issues, VOA3R aims to develop a metadata AP (the VOA3R AP) based on the feedback received from the content providers of the project [2],[3]. This profile will be used for the integration of content within the project and will remain available for future use by additional content providers which wish to make their collections available through the VOA3R federation and platform.

While the content providers of the VOA3R project with existing repositories need to map their metadata AP to the VOA3R AP, the content providers with new repositories can adopt the VOA3R AP and procedures for quality assurance from the very beginning. In the case of the content providers with existing repositories, an enrichment of the metadata description, in order to meet the criteria of the VOA3R metadata application profile might be necessary [4],[5]. As regards this enrichment, there are two options:

- To revise or to extend their internal metadata schema by adding content to their repository or

- To modify their existing metadata application profile, in order for it to be compliant with the VOA3R one.

The content provider will be free to use the option that suits best its collection's needs. It should be emphasized that both options are supported by the VOA3R project.

2.2 The VOA3R Collection

The VOA3R federation consists of content integrated from the following institutional repositories:

2.2.1 Epsilon - Swedish University of Agriculture Sciences (SLU), Sweden. SLU has national responsibility for research and education in veterinary medicine, forestry, and landscape planning. SLU will be provider of metadata to the VOA3R specifications as well as content provider through the Epsilon Open Archive, which contains a significant number of research documents in the form of theses (≈3700), papers and publications (≈500) and other documents (≈550).

2.2.2 OceanDocs - Hasselt University (UHASSELT), Belgium. Hasselt University will provide metadata and data from two different sources: OceanDocs (≈9000 resources) and the Center for Environmental Sciences (≈150 resources), as well as repository development (based on the OceanDocs experiences), and also at a more limited level in the development of tools.

2.2.3 Organic Eprints - International Center for Research in Organic Food Systems (ICROFS), Denmark. ICROFS is a "centre without walls", facilitating the multidisciplinary research in organic food and farming across institutes. It is hosted by the Faculty of Agriculture of the University of Aarhus (DJF). Its main purpose is to initiate, coordinate and monitor high quality organic research, and to stimulate international and transnational research. ICROFS hosts and administrates Organic Eprints (≈6900 resources), the open-access archive on organic research (www.orgprints.com), which will be used in the VOA3R project.

2.2.4 ProdINRA - French National Institute for Agricultural Research (INRA), France. INRA is the largest agricultural research institute in Europe. INRA has signed the Berlin Declaration on Open Access in July 2005. An institutional open archive called ProdInra has been set up to make available the INRA research, which will contribute about 3800 publications (e.g. articles, reports, papers and books) to the VOA3R Repository. Since October 2006, ProdINRA is used by all researchers at INRA.

2.2.5 U-GOV - Consorzio Interuniversitario (CINECA), Italy. CINECA is a Consortium consisting of 36 Italian Universities. Its institutional mission is to support research conducted by the Italian scientific community through supercomputing and its applications. Cineca will participate on VOA3R as content and metadata provider through the U-GOV Repository, offering about 45000 scholarly resources to the VOA3R repository.

2.2.6 ARI Repository - Agricultural Research Institute (ARI), Cyprus. The Agricultural Research Institute (ARI) was established in 1962 and is one of the Departments of the Ministry of Agriculture, Natural Resources and Environment. ARI

undertakes applied and basic research within the wider domain of plant and animal production. Its mission is to provide high quality scientific research using methods that are financially, environmentally and socially sustainable. ARI repository includes full-text publications (in PDF format) from 1965-2007, with the metadata description in FAO AGRIS AP.

More information about the aforementioned institutional repositories and their corresponding contribution to the VOA3R network can be found in Table 2.

3 Generic Process

3.1 Aim of the process

The VOA3R project has developed a methodology for the integration of content, the schema is presented in Figure 1, in order to meet the specific needs of the VOA3R project. The methodology proposed by the VOA3R project consists of four phases, during which the VOA3R content repository will be populated. The Initial Phase will be a Testing Phase which may be considered as a kind of pre-phase, followed by other three phases. All these Phases include a series of generic steps that are proposed to be followed in all phases (the testing phase and the three phases) of the content population and are explained in the Table 1 below.

Table 1. The Testing Phase together with the Three Phases.

Phase	Date	Activities
Controlled testing phase	Jul – Aug ’11	Enrichment of test metadata records using Confolio
Phase 1 of Content Population	Sep – Dec ’11	Integrating the agriculture and aquaculture repositories using the common OAI-PMH specification
Phase 2 of Content Population	Jan– Aug ’12	Integration of repositories without OAI-PMH support and further contents from internal partners
Phase 3 of Content Population	Sep ’12 – May ’13	Enriching the VOA3R portal by external partners and contents

Table 2. VOA3R Content Providers and information on their collections.

Partner name	Swedish University of Agriculture Sciences (SLU)	International Centre for Research in Organic Food Systems (ICROFS)	Consorzio Interuniversitario (CINECA)	Hasselt University (UHASSELT)	French National Institute for Agricultural Research (INRA)	Agricultural Research Institute (ARI)
Collection / Repository Name	Epsilon Repository	Organic Eprints	U-GOV Repository	OceanDocs Repository	ProdINRA Repository	ARI Repository
Resource type	~3700 Theses ~500 Papers and publications ~550 Reports/Others	~10000 Papers ~1400 Organizational resources ~350 Others	~45000 Papers, publications and other types ~1700 News items	~9000 Papers and Reports	~4300 Papers ~340 Books, book chapters & working papers ~ 240 Reports ~ 430 Other documents	~315 Monographies
Content language(s)	Swedish and English	English, German and Danish	Italian, English	English, French, Spanish, Portuguese, Russian, Arabic	French & English	English
Metadata language(s)	There is a "main" language for the metadata record which is the language of the publication.	English	Italian, English	English	French & English	English, Greek
Topics covered by the content	Horticulture, Agricultural Science, Natural Resources, Forest Sciences	Research in Organic food and farming systems	Agriculture, rural development; agricultural economics, management, agribusiness, ICT, information systems, e-business, social economy & rural sociology	Oceanography, Water, Fisheries, Environment, Aquatic science	Agriculture, Environment, Food	Agriculture

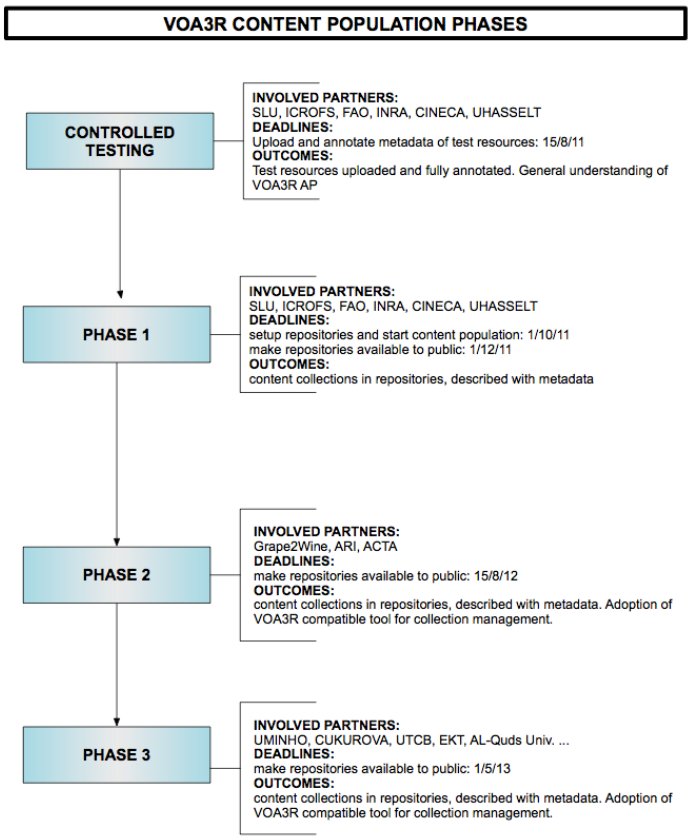


Fig. 1. The four phases of the VOA3R content integration procedure.

3.2 Overview of the process

The proposed four steps are the following:

3.2.1 Uploading/Integration. This step includes the uploading of the resource or the link pointing to this resource. In Figure 2 is presented the uploading/integration process. According to our methodology, a pre-check of the content should be conducted before the uploading takes place, in order to assure that the following basic criteria are fulfilled, by the statement of the content provider:

- the resource does not contain any violent, pornographic or racist content;
- the resource is relevant to the repository it should be inserted (e.g. Agriculture & Aquaculture in the case of VOA3R project) and
- the Intellectual Property Rights do not prohibit the resource to be distributed.

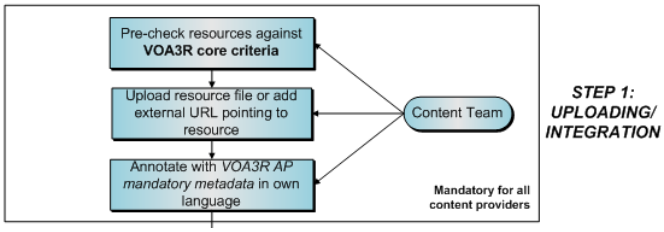


Fig. 2. Step 1 of the VOA3R content integration process.

3.2.2 Enrichment. The Enrichment step, presented in Figure 3, includes the annotation and/or enrichment of the metadata elements, according to the VOA3R set of recommended metadata. The main language of the metadata should be English, while an additional translation of these metadata fields should take place.

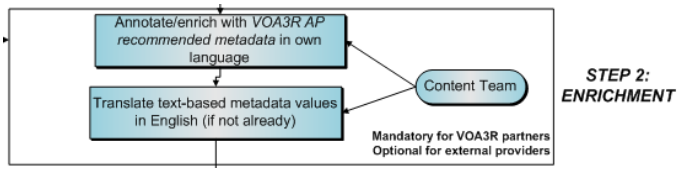


Fig. 3. Step 2 of the VOA3R content integration process.

3.2.3 Validation. Before the publication of the resources, a validation, presented below in Figure 4, of the metadata is necessary. Both the metadata records and the resource are checked against the VOA3R core criteria mentioned above. In case the validation is successful, the metadata records can be published on the VOA3R web portal. The check against the VOA3R core criteria in this step is not done by the content team member but by the content validator.

If the tool used by the partners that are willing to offer their content is not one of the tools already used by the VOA3R federation, it should have the ability to expose metadata through the OAI-PMH standard and map these metadata to the VOA3R application profile. The exposure of the metadata through the aforementioned standard should be done in this step.

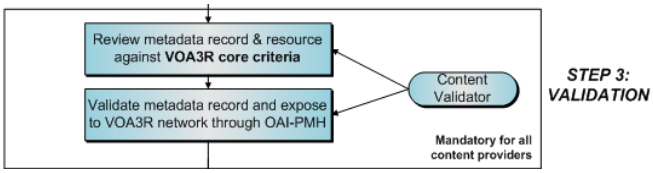


Fig. 4. Step 3 of the VOA3R content integration process.

3.2.4 Quality Review/Assessment. After the resources are published, respecting the schema from Figure 5, a number of them will be evaluated in terms of the quality of

both their content and their metadata. This evaluation will be performed with the use of predefined check grids in order to review the quality of the content and the metadata, as well as their relevance with the content.

According to the proposed methodology, the quality of the metadata is reviewed manually by a peer-reviewer (an expert in the field of the resource belonging to one of the participating organizations which is). The internal quality review schema will involve the project partners [2], who will cross-check each other's metadata records without having the permissions to change it. As regards the completeness of the metadata description, an automatic mechanism will check this aspect and will warn in case the mandatory fields of the metadata of any resource are not filled in.

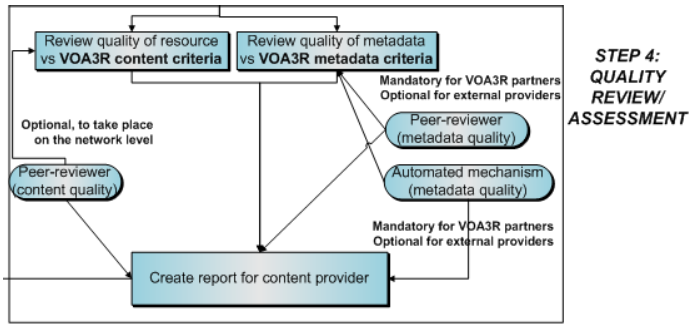


Fig. 5. Step 4 of the VOA3R content integration process.

The aforementioned four steps are for all the content providers, using any system as a Content Management System to connect to VOA3R repository.

4 Scenario of use: The Testing Phase

In this Section we will discuss only the Testing Phase, as the other three Phases could be subject for another scientific paper. The Testing Phase includes all steps mentioned in the previous Section. The testing Phase it is addressed to all the content providers using Confolio. Confolio (http://voa3r.confolio.org) in Figure 6 below, is a repository tool, which is used for creating, storing, indexing and retrieving the metadata description for each resource which is stored in a networked repository.

Fig.6. Confolio platform.

During the testing phase, the following steps will take place:

1. Uploading/Integration:
   Users will be asked to register in the VOA3R Confolio tool (http://voa3r.confolio.org) presented in Figure 7 below. After they successfully log in, they will have to upload a number of digital resources, respecting the VOA3R Core Criteria Grid and provide the mandatory metadata description for each one of the resources.

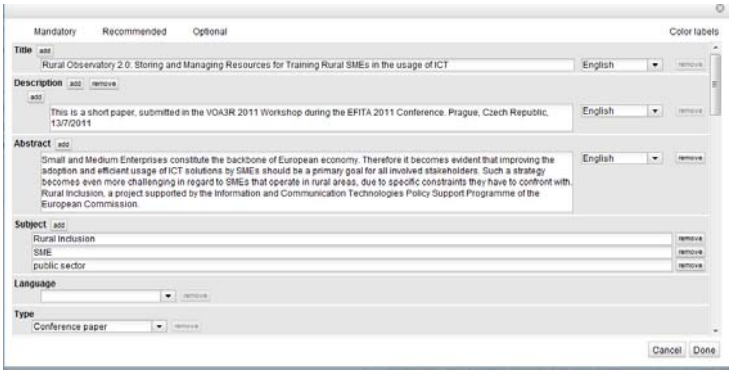


Fig.7. The Uploading/Integration step in the Confolio platform.

2. Enrichment:
   Partners will be asked to provide the recommended metadata description as well and translate in English all the metadata fields used. The user interface will be like the one presented in Figure 8.

Fig. 8. The Enrichment step in the Confolio platform.

3. Validation:
   Uploaded resources and their metadata will be checked against the VOA3R Core Criteria and they will be published. The grid with the VOA3R Core Criteria that should be respected is presented in Figure 9.

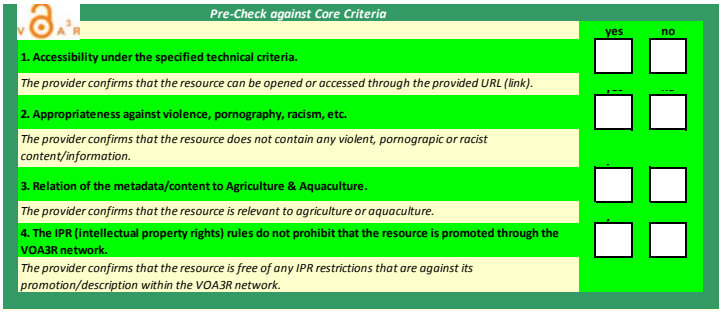
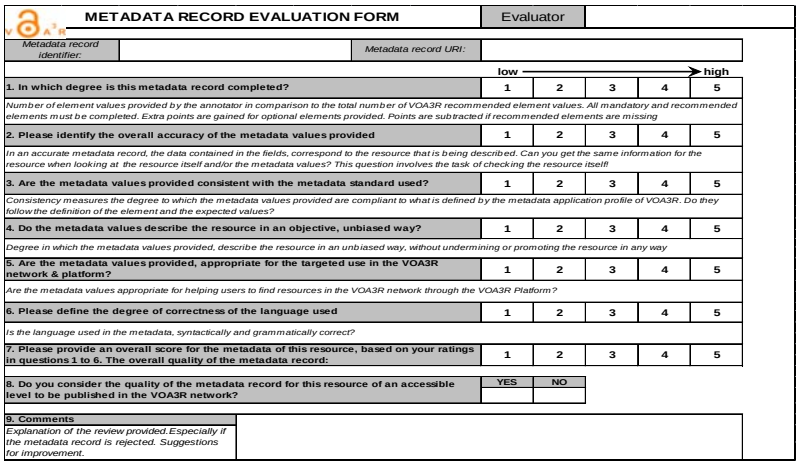


Fig.9. The VOA3R Core Criteria.

4. Quality review/assessment:
   Uploaded resources will be reviewed [4], according to the evaluation form Figure 10, in terms of the content and the use of metadata by the WP5 leaders and a report will be sent back to the content teams for revising the metadata description.



METADATA RECORD EVALUATION FORM						Evaluator	
Metadata record identifier:		Metadata record URI:					
			low →				high
1. In which degree is this metadata record completed?			1	2	3	4	5
Number of element values provided by the annotator in comparison to the total number of VOA3R recommended element values. All mandatory and recommended elements must be completed. Extra points are gained for optional elements provided. Points are subtracted if recommended elements are missing							
2. Please identify the overall accuracy of the metadata values provided			1	2	3	4	5
In an accurate metadata record, the data contained in the fields, correspond to the resource that is being described. Can you get the same information for the resource when looking at the resource itself and/or the metadata values? This question involves the task of checking the resource itself!							
3. Are the metadata values provided consistent with the metadata standard used?			1	2	3	4	5
Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata application profile of VOA3R. Do they follow the definition of the element and the expected values?							
4. Do the metadata values describe the resource in an objective, unbiased way?			1	2	3	4	5
Degree in which the metadata values provided, describe the resource in an unbiased way, without undermining or promoting the resource in any way							
5. Are the metadata values provided, appropriate for the targeted use in the VOA3R network & platform?			1	2	3	4	5
Are the metadata values appropriate for helping users to find resources in the VOA3R network through the VOA3R Platform?							
6. Please define the degree of correctness of the language used			1	2	3	4	5
Is the language used in the metadata, syntactically and grammatically correct?							
7. Please provide an overall score for the metadata of this resource, based on your ratings in questions 1 to 6. The overall quality of the metadata record:			1	2	3	4	5
8. Do you consider the quality of the metadata record for this resource of an accessible level to be published in the VOA3R network?			YES	NO			
9. Comments							
Explanation of the review provided.Especially if the metadata record is rejected. Suggestions for improvement.							

Fig. 10. The metadata evaluation form.

The Testing Phase will take place in three periods:

- The first period is a trial annotation period when all the content providers will upload on Confolio a small number of sample resources in order to experiment and to get familiar with the proposed elements of the VOA3R metadata AP, as well as with the proposed value sets. At this current stage of the Testing Phase the annotations are only in English. The samples which will be uploaded are part of the promised content for the repository. The scope of this first period is to verify the design choices at an early stage and the feedback of the content providers is welcome and it will help in improving the process.
- In the second period, the common scheme and the metadata enrichment will be revised. The content providers that will adopt the common scheme will develop internally the appropriate infrastructure to make available the VOA3R profile through OAI-PMH protocol. The content providers that will not adopt the VOAR profile will need to add extra annotations through Confolio. If in the first period there were tested a few samples, in this second period, there will be more samples. Moreover the annotations will not be only in English but in multiple languages.
- In the third period, final adjustments to the common metadata schema will be performed. Furthermore the final version in XML and RDF formats will be released. Content providers will focus their efforts on populating the content of their repositories with additional metadata. The content providers that will adopt the VOA3R metadata schema will have to populate the new elements using their own tools and procedures while the content providers that will not extend their metadata structure must provide with the additional metadata.

5 Conclusions

The need for working on standardized methodologies and procedures in the area of content integration from various repositories is constantly rising. What this paper has innovative is the process of integration of different repositories from diverse users, users with their own system or users with a system compatible with ours. Therefore this paper provides a methodology for content integration from various agricultural repositories to a single network, by suggesting a uniform method for the different phases of the process, as well as a common metadata AP to be used by all partners who integrate their content in the VOA3R collection. We would like to emphasize that the access to the metadata of the content will be open. Thus we are trying to head towards the trend of the society, open access. Regarding the content we cannot guarantee open access because of the copyrights issues.

Acknowledgments. The work presented in this paper was conducted under the framework of the VOA3R project funded by the European Commission through the ICT PSP Programme (ICT PSP), Theme 4 - Open access to scientific information. All authors have been supported with funding by GRNET during the implementation of this work, in the context of the above mentioned EU project.

References

1. Organic.Edunet: Deliverable D4.1, Methodology for Content Population. A Multilingual Federation of Learning Repositories with Quality Content for the Awareness and Education of European Youth about Organic Agriculture and Agroecology (2009)
2. Patrikakis C.Z., Koukouli M., Papadopoulos G. K., Sideridis A. B.: Evaluating Behavioral Change in Multigroup Collaboration for Content Publishing Over the Web. Social Science Computer Review 2009; 27; 59 originally published online Jun 5, 2008 DOI: 10.1177/0894439308319449
3. Evaluating Behavioral Change in Multigroup Collaboration for Content Publishing Over the Web, http://ssc.sagepub.com/cgi/content/refs/27/1/59, Accessed 26 June 2011
4. Palavitsinis N., Manouselis N., Sanchez S.: Evaluation of a Metadata Application Profile for Learning Resources on Organic Agriculture, in Proc. of 3rd International Conference on Metadata and Semantics Research (MTSR09), Milan, Italy (2009)
5. Palavitsinis N., Ebner H., Manouselis N., Sanchez S., Naeve A.: Evaluating Metadata Application Profiles Based on Usage Data, in Proc. of the International Conference on Digital Libraries and the Semantic Web (ICSD 2009), Trento, Italy (2009)