ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

# Annotating Experimental Records Using Ontologies

Alexander Garcia[1], Olga Giraldo[2], Jael Garcia[3]

[1]University of Arkansas, Biomedical Informatics, Medical Center, Little Rock, Arkansas, USA
[2]National University of Colombia in Palmira, Valle, Colombia
[3]Universität der Bundeswehr, Munich, Germany

**Abstract.** By combining a complex mixture of electronic and paper-based records, researchers carefully document their daily research activities. Although managing such a mixture is a common practice, much information recorded in laboratory notebooks in this manner is often lost for practical purposes. Moreover, lab notebooks are usually disconnected from other information resources that researchers frequently use. Interestingly, although Electronic Notebooks are available, these have not been widely adopted. Here we present our approach to the problem of managing knowledge in Electronic Laboratory Notebooks. We combine elements from the Semantic Web, e.g. ontologies supporting organization and classification, with elements from Social Tagging Availability: www.biotea.ws.

## 1 Introduction

Here we present a knowledge-based approach to managing laboratory information; it combines elements from the Semantic Web (SW), e.g. ontologies supporting organization and classification, with elements from Social Tagging Systems, e.g. collaboration. We have developed several ontologies supporting the annotation of experimental data for some of the processes routinely run at the Center for International Tropical Agriculture (CIAT) biotechnology laboratory. To identify those processes and practices we analyzed 15 laboratory notebooks together with their corresponding electronic records, e.g. XLS files. We identified data types, metadata, organization, retrieval and sharing strategies, sources of data and information, as well as ontology support for the annotation of laboratory information. Central to our approach is the symbiosis between ontologies and social tagging systems [1, 2]. As ontologies do not fully cover the whole domain, the annotation of laboratory notebooks is achieved by combining ontology-based and user-generated tags; generating in this manner a social network built upon tagged concepts. Our scenarios range from basic laboratory techniques such as PCR, DNA extraction, Electrophoresis, and others, to those involving complex biological phenotype-genotype relationships. We are extending and reusing existing ontologies such as the Ontology for Biomedical Investigations (OBI) [3], the Chemical Entities of Biological Interest ontology (ChEBI) [4], Plant Ontology (PO) [5], Gene Ontology (GO) [6], Annotation Ontology [7], amongst others.

## 2 Experimental Records as Knowledge Repositories

For the analysis of laboratory practices and notebooks we closely followed the methodologies proposed by Tabard *et al.* [8] and Garcia et al [9]. We interviewed 10 biologists, analyzed their laboratory notebooks, 15 in total, and electronic records; our field observation went on for a period of six months. Several interviews were held between the elicitor and the researchers; use cases representing processes, practice and involved data were constantly being built; these uses cases were the basis for our ontology development as well as for our iterative prototyping.

Researchers pay little attention to the sequence of the information; for instance, notes for long experiments are spread throughout the entire laboratory notebook and stored in several electronic files. Researchers tag in order to establish an organization strategy; however, the tagging strategy is very personal. They also add marginal notes; these were usually comprised of few descriptive words located in visible areas of the corresponding page. It was also observed that the vocabulary used to tag was significantly overlapping amongst researchers who were working on conceptually closer topics; this trend has previously been reported by Marlow *et al.* [10]. For instance, researchers studying genes involved in drought tolerance share information with those who participate in field studies involving those samples that were genetically modified to be more tolerant to the lack of humidity and water. Researchers also deal with electronic records; they store photos, XLS files, outputs of specialized analysis software, etc. Researchers tend to store and manage their files in their PCs; electronic records also come from LIMSs. The rhetorical structure, and the ontologies related to the components of such structure, is presented in Figure 1.

## 3 Our Approach: Towards Self-Descriptive Documents

Interestingly, the lack of strategies for organizing and managing knowledge in documents (paper-based and electronic files) had previously been reported, albeit in a different domain, by Paganelli *et al.* [11]. Semantic annotation of features facilitates the self-descriptiveness of documents; the availability of such semantics is key when managing organizational knowledge [12]. Documents should be able to "know about" their own content for automated processes to "know what to do" with them. By delivering ontology-based annotation facilities combined with tagging functionalities researchers are adding that descriptive layer in order to i) speed up information retrieval, ii) facilitate collaboration iii) generate an organization strategy – sometimes mainly understood by the laboratory notebook owner.

We have structured the descriptive layers by reusing and extending existing ontologies. For supporting the annotation within our scenario we have identified three main layers, namely: i) that related to the document itself, ii) the annotation layer, and iii) that related to the experiment. For the document we investigated several metadata standards such as Dublin Core (DC) [13], AgMes [14], AGROVOC and the National Cancer Institute thesauri (NCIt); annotations were structured by means of the AO [7]; experimental information was structured by reusing and extending biomedical ontologies such as OBI, ChEBI, AGROVOC, PO, and GO. An illustration of our layered approach is presented in Figure 1 and Figure 2.

The AO is structuring the semantic annotation as well as the tags generated by users. In this way we are supporting complex SPARQL queries involving several ontologies, for instance: "*retrieve from the eLabBook the pages tagged by Tim Andrews or Lisa Watson with the tags rice and iron for which there is a LIMS data entry*". This query involves highly interrelated information, covering aspects related to the pages within the ELN (document), annotation and experimental information.

```
SELECT ?eLabBook ?page
WHERE {
 ?annotation ann:annotates ?page .
 ?annotation  pav:createdBy  ?user  .  ?user
foaf:name ?userName .
 FILTER(?userName   =   “lisa watson”   ||
?userName = “tim.andrews”).
 ?annotation ao:hasTag ?tag .
 ?tag tags:name ?tagName .
 FILTER(?tagName  =  "rice"  ||  ?tagName  =
"iron")
 ?eLabBook hasPage ?page .
 ?page hasLIMSDataEntry ?lims
}
```

Our document layer provides the ontology for describing the laboratory notebook as well as classes and properties for representing relations with resources such as the LIMS. It has concepts such as "*creator*" (DC), "*investigator*" (NCIt) and "*laboratory procedures*" (NCIt). As researchers store information as they produce it, time is an issue. For instance, researchers may start to record the growth of a plant at intervals of 15 days; this usually also involves taking pictures of the plant, sampling the soil and keeping daily records for atmospheric conditions. The records for his/her observations will be spread all over the laboratory notebook, making it difficult for the researcher to have a unified view of the work. For such situations time stamps are not sufficient; the property "*has_labprocedure*" is practical because this property can be further specialized by "*laboratory procedures*" (NCIt) and "*study subject role*" (OBI_0000097); facilitating thus the interlinking of records so that researchers may retrieve unified views for specific "*laboratory procedures"* combined with "*study subject roles*" – e.g. "*plant structure*" (PO:0009011).

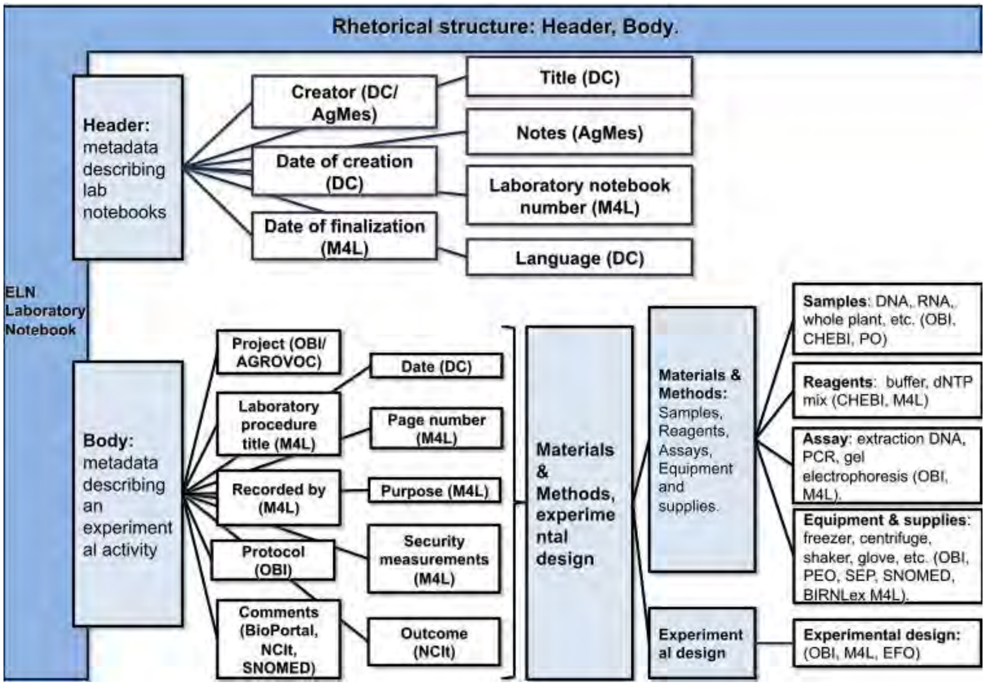
Rhetorical structure: Header, Body.
Header: metadata describing lab notebooks
Creator (DC/ AgMes)
Date of creation (DC)
Date of finalization (M4L)
Title (DC)
Notes (AgMes)
Laboratory notebook number (M4L)
Language (DC)
ELN Laboratory Notebook
Body: metadata describing an experimental activity
Project (OBI/ AGROVOC)
Laboratory procedure title (M4L)
Recorded by (M4L)
Protocol (OBI)
Comments (BioPortal, NCIt, SNOMED)
Date (DC)
Page number (M4L)
Purpose (M4L)
Security measurements (M4L)
Outcome (NCIt)
Materials & Methods, experimental design
Materials & Methods: Samples, Reagents, Assays, Equipment and supplies.
Experimental design
Samples: DNA, RNA, whole plant, etc. (OBI, CHEBI, PO)
Reagents: buffer, dNTP mix (CHEBI, M4L)
Assay: extraction DNA, PCR, gel electrophoresis (OBI, M4L).
Equipment & supplies: freezer, centrifuge, shaker, glove, etc. (OBI, PEO, SEP, SNOMED, BIRNLex M4L).
Experimental design: (OBI, M4L, EFO)

**Figure 1.** The rhetorical structure

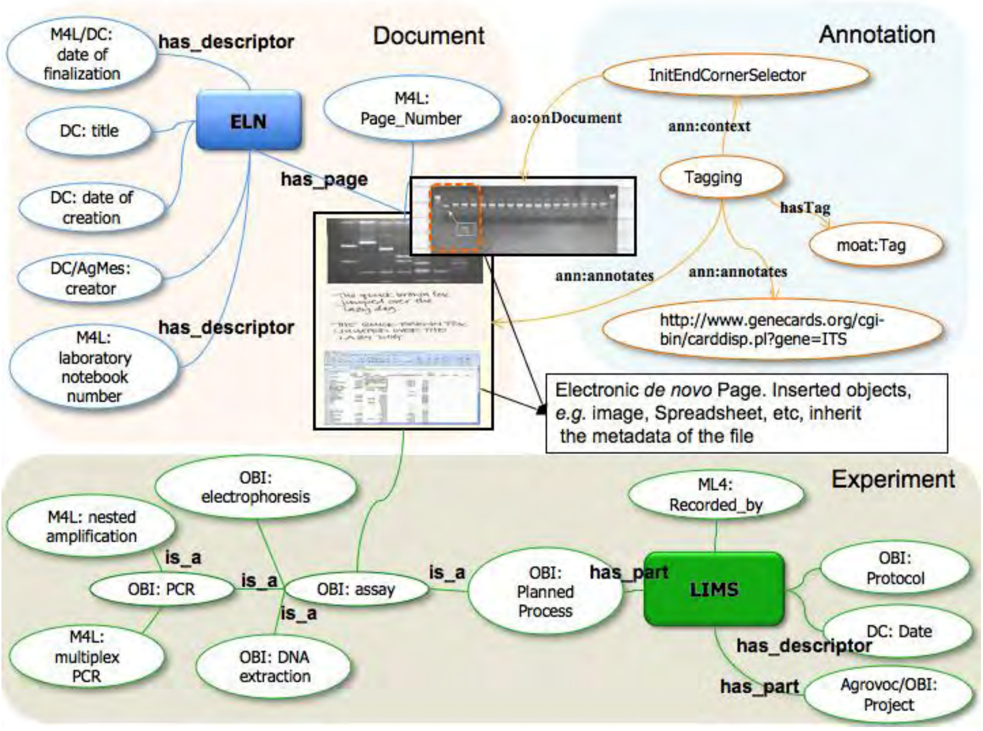

**Figure 2.** Our layered approach[1]

[1] M4L stands for Metadata for Laboratory; it denotes those ontologies developed specifically for our scenarios. M4L ontologies are available at www.biotea.ws

### 3.1 Structuring the Annotations

For structuring the tagging we decided to use the AO. It was conceived to support the annotation of documents; it can be used to save the tagging data and publish it as Linked Data. Tags are part of the organizational strategy used by researchers in their lab-notebooks; tags are also used to relate specific areas, within pages of the lab-notebooks, to internal and/or external resources. As illustrated in Figure 3, not only is it possible to annotate the entire ELN page, but also a selected portion of it; the AO facilitates the definition of relations between electronic pages and internal/external resources. Also, as part of the example presented in Figure 3, there are two annotations made by the same user; being the user represented by her/his FOAF. Both ANNOT_1 and ANNOT_2 are Qualifier annotations, *i.e.* an ontological term – *GeneBank:AB005238*, is attached to the tag – *Partial sequence on psy promoter*. ANNOT_1 annotates a portion of the image in the ELN and relates it to an ontological term, which is also related to a scientific paper by means of ANNOT_2. In this way it is possible to enrich the information in the ELN with ontological terms, free text, papers, images, videos, and anything for which there is an URI. Having other type of annotations such as Note, Definition, and Erratum is also possible.

By tagging laboratory notebooks researchers are generating clouds of tags. As laboratory notebooks don't have tables of contents, users identified the clouds of tags as a valuable resource for rapid inspection of contents. By facilitating the generation of tags, combining those coming from ontologies with those provided by users, we are supporting queries such as "*retrieve from the eLabBook those pages having an EXCEL spreadsheet and that have been tagged by Tim Andrews with the tag rice and optionally with PCR*".

```
   SELECT ?eLabBook ?page ?file
   WHERE {
       ?annotation ann:annotates ?page .
       ?annotation pav:createdBy <http://www
.tags4lab.org/foaf.rdf#tim.andrews> .
       ?annotation ao:hasTag ?tag .
       ?tag tags:name "rice" .
       OPTIONAL
{?tag moat:tagMeaning OBI:PCR} .
       ?eLabBook hasPage ?page .
       ?page hasExcel ?file
   }
```

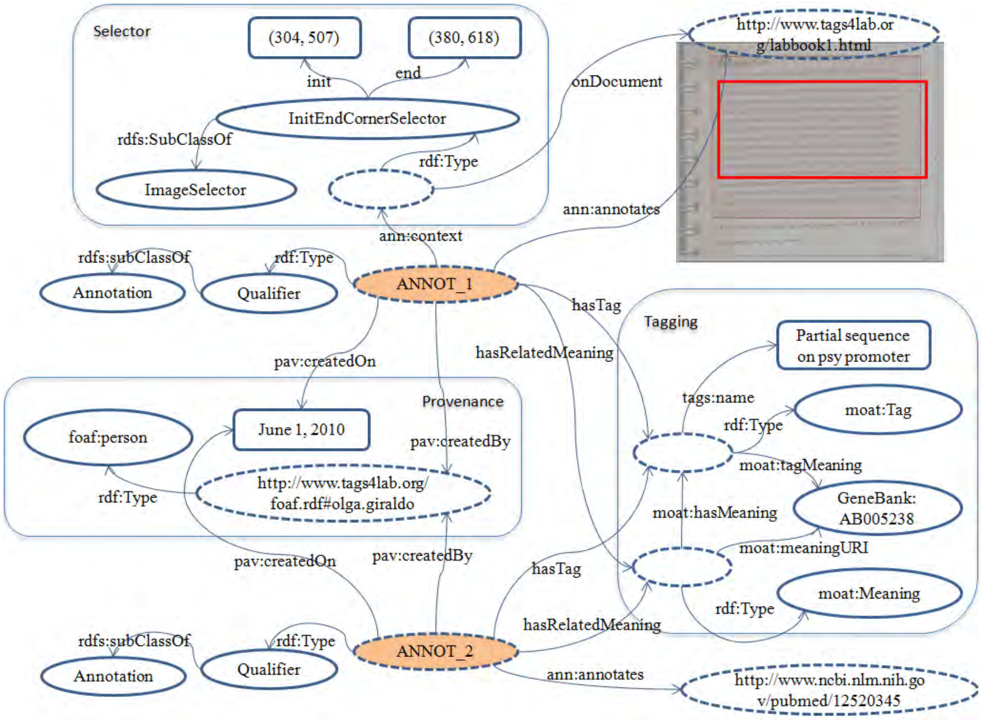

**Figure 3.** Supporting the annotation of laboratory records with AO

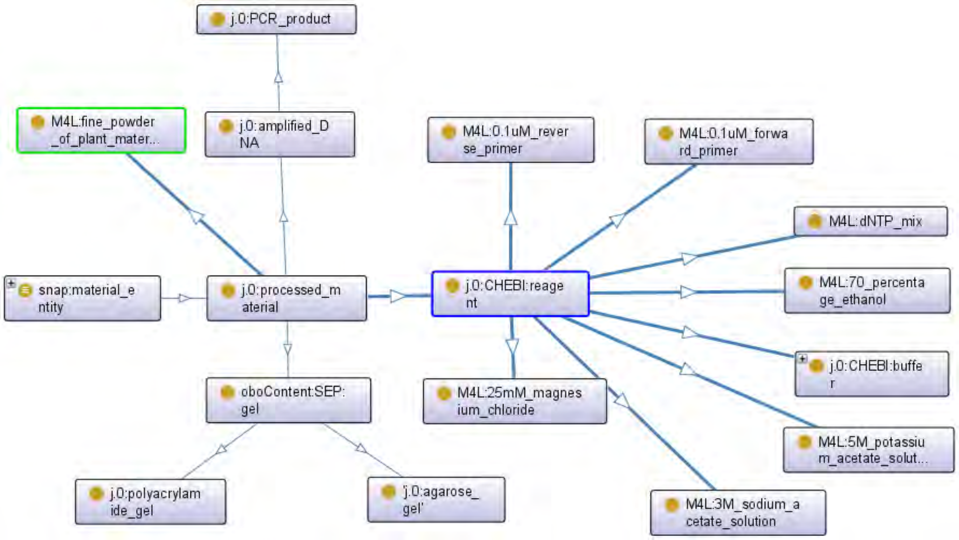

**Figure 4.** A snapshot of M4L. j.0=undeclared name space.

## 3.2 Experimental Data

Describing biomedical investigations involves bringing together a wide variety of highly interrelated data. There are various levels of complexity and granularity, combined with a wide range of materials and equipment [15]. Within the structure provided by the Basic Formal Ontology [16] OBI defines an extendible set of terms aiming to describe biological and clinical investigations. For the experimental layer we have followed the structure provided by BFO and OBI (version 2009-11-06, aka version 1.0). OBI defines an "investigation" (OBI_0000066) as a "process" (BFO) that involves, amongst others, the general planning of the "study design" (OBI_0500000), its corresponding execution, the documentation of results and the "interpreting data" (OBI_0000338) so that conclusions can be derived and supported.

Similarly to the clinical domain, processes in plant biotechnology also involve several sub-processes. From BFO, OBI uses "*material entity*" (BFO) as the basis for physical artifacts. Material entity is "*an independent continuant that is spatially extended whose identity is independent of that of other entities and can be maintained through time*". A *material entity* is, for instance, a collection of random bacteria, a chair, or the dorsal surface of the body. Material entities can have "*roles*" (BFO); for instance the "*study subject role*" (OBI_0000097). "*Functions*" depend on the design or physical structure of the entity; for instance, "*measure function*" (OBI_0000453), "*freeze function*" (OBI_0000375). Functions are considered inhered, "*inheres in*" (BFO) by the material entity and "*realized by"* (BFO) the "*role"* that is assumed by the "*material entity*" within the process.

As illustrated in Figure 4, we are reusing and extending OBI in combination with other ontologies so that our use cases are fully covered; these have been selected from those laboratory procedures, "*assay"* (OBI_0000070), commonly carried out by the Biotechnology group at CIAT; a snapshot of M4L is presented in Figure 4. To illustrate some of the experimental ontologies that have been developed we selected the small scale extraction of high quality DNA "assay" (OBI_0000070). Three planned processes are part of this assay, namely: harvesting the plant material, pulverizing it, and extracting the DNA.

## 3.2 Sample Preparation for Assay and DNA Extraction

Both, harvesting the plant material and pulverizing it, illustrate the "*sample preparation for assay*" (OBI_0000073) class from OBI. Initially researchers use "*scissors*" (SNOMED-CT ID 64973003), for which there is a "*mechanical function"* (OBI_0000379) to obtain the "*juvenile leaf*" (PO:0006339) or an "*adult leaf*" (PO:0006340). The "*leaf*"

(PO:0009025) assumes the “*study subject role*” (OBI_0000097). The “*leaf*” is stored in a “*reclosable bag*” (M4L) that is stored in a “*portable ice chest*” (M4L). The vegetal material is then stored in a “*freezer”* (PEO, “*freeze function*” OBI_0000375). Pulverizing the “*leaf*” starts by taking the “*leaf*” out of the “*freezer”*; the “*frozen leaf*” (M4L) is a “*material entity*” (snap:MaterialEntity). This material is then converted into a “*fine powder*” (M4L); such “*fine powder*” is a “*processed material*” (OBI_0000047).

A “*microcentrifugue tube*” (M4L) is then used to store the “*fine powder*” (M4L). “*DNA extraction buffer*” (M4L), the buffer’s “*role*” is to dissolve the tissue; facilitating in this manner the extraction of the DNA. The whole process, from the “*fine powder”* is illustrated in Figure 5. The DNA extraction ontology has over 140 classes; we are reusing BFO properties such as “inheres in” (BFO_0000052), “bearer of” (BFO_0000053), “realized by” (BFO_0000054), and “realizes” (BFO_0000055). Other sources for relationships come from the Relation Ontology [17]. Twenty-nine fully documented new terms, from M4L, have been added to the OBI structure that we are reusing; terms from other ontologies are also being reused.

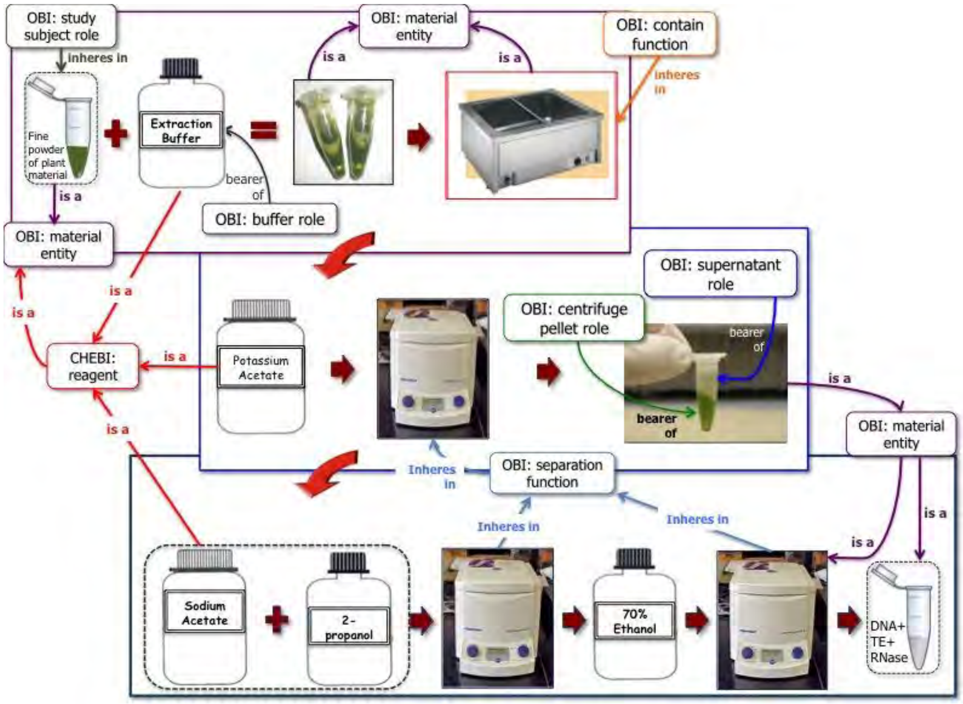

**Figure 5.** Extracting DNA

## 4 Discussion

Biologists have so far managed to balance and combine paper-based records with an intricate network of electronic records (LIMS, spreadsheets, URLs, etc). However, as it is necessary to process more and more data in a systematic interoperable manner, the information management infrastructure will have to facilitate data capture/processing in an integrative way; more importantly, the semantic layer within these infrastructures needs to be vastly improved. Such layer requires interoperable, well modularized ontologies rather than monolithic ones. Although several ELNs have been proposed [18, 19], and replacing paper-based records has been a consistent trend for several years, the technology has not yet been widely adopted [18]; Laboratory Information Management Systems (LIMS) in combination with paper-based laboratory notebooks continue to be commonly used; particularly in academic environments [8, 20].

Sharing and organizing information happens on a concept basis; for instance, researchers studying genes involved in iron transport share information with those who undertake nutritional studies assessing the effects of iron intake in human populations. Such concept-based folksonomy was easily observed; ontologies supporting the annotation of laboratory records and practices made it easier for researchers to share and interact. By the same token, being able to tag with user generated tags as well as ontology-based tags, facilitated the organization of information. Interestingly, although tagging practices were personal, these were similar amongst those researchers working on conceptually similar projects. Tags were also a valuable resource providing new terms for our ontologies.

## 5 Conclusions and Future Work

Using ontologies to support the annotation of experimental activities requires highly interoperable ontologies. Although there is a generic extensible ontology for relations in the biomedical domain, not all of them are actively using it. Also, although biomedical ontologies are pursuing a thoughtful and important ontology development standardization effort; there are still methodological gaps. OBI facilitates the description of biomedical experiments; such effort implies interoperating with other ontologies; from our experience interoperability between OBI and OBO ontologies was not a straightforward process. Standardization efforts based on minimal amounts of information, grounded in existing ontologies, are important for facilitating interoperability across laboratories; such efforts should focus on providing easily implementable data capture templates.

We have presented a semantic layered approach that facilitates the self-description of experimental records from the Biotechnology laboratory at CIAT. We have reused CHEBI, OBI, BIRNLex as well as other ontologies; within our OBI scaffold we have added 145 terms, all of them extracted from our experimental records. We envision a paperless laboratory in which Ubiquitous Computing takes advantage of SW technology, for supporting knowledge management, and folksonomy principles for facilitating the collaboration. We have started by making extensive use of ontologies for supporting knowledge management, by the same token we are facilitating interaction in similar ways to those currently available in social networks; our interaction is based upon research activities and concepts. In the near future we will improve the usability of our prototype, we are also planning to release the software to the open source community; we are currently continuing with our ontology development effort.

## References

1. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering. International World Wide Web Conference – Workshop on Social and Collaborative Construction of Structured Knowledge (CKC), Canada (2007)
2. Almeida, A., Sotomayor, B., Abaitua, J., López-de-Ipiña, D.: Folk2Onto: Bridging the gap between social tags and ontologies. European Semantic Web Conference, Tenerife, Spain (2008)
3. Brinkman, R., Courtot, M., Derom, D., Fostel, J., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Soldatova, L., Stoeckert, C., Turner, J., Zheng, J., consortium, t.O.: Modeling biomedical experimental processes with OBI. Journal of biomedical semantics **1** (2010) S7
4. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: Chemical entities of biological interest: an update. Nucleic Acids Research (2009)
5. Pankaj, J., Shulamit, A., Katica, I., Elizabeth, A., Kellogg. , Susan, M., Anuradha, P., Leonore, R., Seung, Y., Rhee., Martin M., S., Mary, S., Lincoln, S., Peter, S., Leszek, V., Doreen, W., Felipe, Z.: Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. Comparative and Functional Genomics **6** (2005) 388-397
6. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel, T.L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G.: Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics **25** (2000) 25-29
7. Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T.: An open annotation ontology for science on web 3.0. Journal of biomedical semantics **2** (2011) S4
8. Tabard, A., Eastmond, E., Mackay, E.W.: From Individual to Collaborative: The Evolution of

Prism, a Hybrid Laboratory Notebook. Computer Supported Cooperative Work. ACM, San Diego, California, USA (2008)

9. Garcia, A., O'Neill, K., Garcia, L.J., Lord, P., Stevens, R., Corcho, O., Gibson, F.: Developing Ontologies within Decentralised Settings. In: Chen, H., Wang, Y., Cheung, K.-H. (eds.): Semantic e-Science, Vol. 11. Springer US (2010) 99-139

10. Marlow, C., Naaman, M., Boyd, D., Davis, D.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. 7th Conf. Hypertext and Hypermedia. ACM Press, Edinburgh, Scotland, United Kingdom (2006) 31-40

11. Paganelli, F., Pettenati, M.C., Giuli, D.: A Metadata-Based Approach for Unstructured Document Management in Organizations. Information Resources Management Journal **19** (2006) 22

12. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics **4(1)** (2005) 15

13. http://dublincore.org: Dublin Core Metadata Initiative (2010)

14. http://aims.fao.org/en/agmes-metadataset: AgMes metadata element set. (2010)

15. Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Soldatova, L.N., Stoeckert Jr., C.J., Turner, J., Zheng, J., consortium., t.O.: Modeling biomedical experimental processes with OBI. Journal of Biomedical Semantics **1** (2010) 11

16. Grenon, P., Smith, B., Goldberg, L.: Biodynamic Ontology: Applying BFO in the Biomedical Domain. Ontologies in Medicine (2004) 20-32

17. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax , J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in Biomedical Ontologies. Genome Biology **6** (2005) R46

18. Van Eikeren, P.: Intelligent Electronic Laboratory Notebooks for Accelerated Organic Process R&D. Organic Process Research & Development **8** (2004)

19. Talbott, T., Peterson, M., Schwidder, J., Myers, J.D.: Adapting the electronic laboratory notebook for the semantic era. In: McQuay, W., Smari, W.W., Kim, S.-Y. (eds.): International Symposium on Collaborative Technologies and Systems. IEEE Computer Society (2005)

20. Butler, D.: Electronic notebooks: A new leaf. Nature **436** (2005) 20-21