Preprint of paper accepted to the ECDL 2007.

Named Entity Identification and Cyberinfrastructure

Alison Babeu, David Bamman, Gregory Crane, Robert Kummer, and Gabriel Weaver

Perseus Project, Tufts University

Abstract. Well-established instruments such as authority files and a growing set of data structures such as CIDOC CRM, FRBRoo, and MODS provide the foundation for emerging, new digital services. While solid, these instruments alone neither capture the essential data on which traditional scholarship depends nor enable the services which we can already identify as fundamental to any eResearch, cyberinfrastructure or virtual research environment for intellectual discourse. This paper describes a general model for primary sources, entities and thematic topics, the gap between this model and emerging infrastructure, and the tasks necessary to bridge it.

1 Introduction

New terms such as eScience and eScholarship have emerged to describe the qualitatively distinctive processes of intellectual life possible in a digital age. We have begun debating what underlying cyberinfrastructure will be necessary to support virtual research environments (VRE) that support scholars in various disciplines [11, 19]. This paper describes work towards a VRE for the humanities in general, with a current pragmatic focus on Greco-Roman studies.

As we explore Greco-Roman research within a digital library we find three general and complementary challenges. First, we need to be able to integrate broad interdisciplinary and deep domain knowledge. Libraries have developed vast, general classification schemes and authority lists to structure all academic knowledge, but humanists, like their colleagues in the sciences, have created their own authority lists that go beyond, and are often unconnected with, library data. Second, we need to be able to customize general services. Google, for example, has produced a service that identifies and maps place names in its digitized books, but this service will not reach its full potential until scholarly communities have integrated into it their expert knowledge about people, places, etc. Third, we need scalable methods to absorb decentralized contributions, large and small. The general public is very good at disambiguating references to people, places, and other entities, as witnessed by the millions of accurate disambiguating links (links between ambiguous names and their articles) in Wikipedia[23].

All three of these problems rely on a single underlying infrastructure: the ability to canonically refer to people, places, and objects in a text, and to integrate knowledge about those entities from disparate sources. Data structures

such as CIDOC CRM[7], FRBRoo[9], and MODS/MADS[17, 16] provide a foundation for this infrastructure, but we must still build on top of them to create useful services. This paper reports on the extent to which we have had to supplement existing resources (data structure, content, and algorithms) as we develop a VRE. We offer a general logical model of the system and then report on the issues that have arisen at three distinct layers within this system. While we focus upon the Greco-Roman cultural heritage that all Europe shares, the specific issues are relevant to other cultural heritage domains and the underlying model supports much intellectual activity beyond the humanities.

2 Semantic Classification and Named Entities

Our work with cultural heritage materials has led us to identify five layers of scholarship, as illustrated in Figure 1. Surrogates in the library include critical editions reconstructing literary texts, documentary editions that reconstruct particular written artifacts such as manuscripts or papyri, archaeological surveys, descriptions of buildings, and catalogues of artifacts. These sources generally strive for transparency, documenting the current evidence and reconstructing the original text or site as it appeared at some point in the past. Secondary sources include reference works, articles and monographs that explore original ideas. Reference works and secondary sources, however, both supplement observational and reconstructive data as reported in library surrogates with direct observation (where places or artifacts survive).

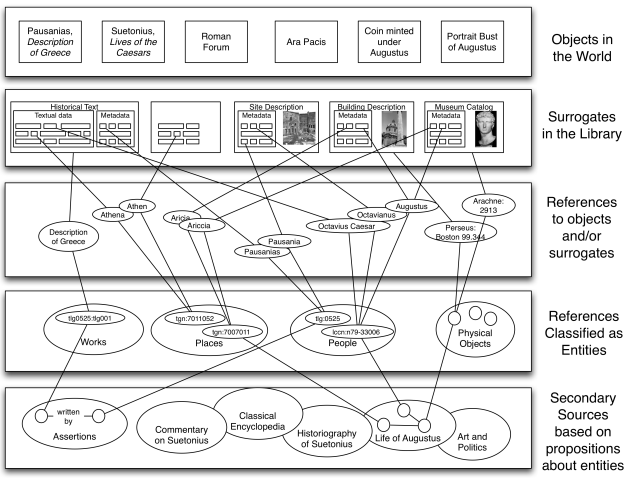

Fig. 1. Model of digital scholarship

Our model emphasizes two layers between surrogates and secondary sources. In the first we find references with some semantic classification - in this level, we

are populating ontologies such as the CIDOC CRM or FRBRoo. In the simplest case, we have subsets of objects: any arbitrary chunk of text extracted for discussion is a quotation; any subset of an object is a detail. Insofar as quotations or details are meaningful, they advance some particular point. Articles of daily life - cooking implements, furniture, jewelry - may form hierarchical classes of object on which we base analysis of social history. In many cases, however, we need to go beyond this kind of semantic classification and associate specific references with entities from the world. We want to associate a reference to a temple of Apollo to a particular archaeological site; to determine not only that Alexandria is a place but which of the many Alexandrias is meant; and to establish whether a given Antonius was the son, grandson, father - or some other person with this name. We are not simply interested in altars but in the monumental Ara Pacis constructed to memorialize the peace which Augustus brought to the Roman world.

In our view, annotating and aggregating references, whether classified by semantic class or as particular entities, are the fundamental processes of scholarship in the humanities – the degree to which an argument bases itself on such annotations defines the degree to which it qualifies as serious intellectual discourse, whether the author holds an endowed chair or is an amateur contributing to Wikipedia. We and others in the field of classics have needed to extend the pre-existing metadata from libraries and general services from information science.

3 Extending Metadata

In twenty years of continuous development, we have attempted, with varying success, to integrate first-class publications of art and archaeological data with first-class textual sources. Most collections will focus on one or the other. Textual collections may include a few illustrative images but lack professional cataloguing information and metadata; while the art and archaeology collections will quote primary sources, often translated and without machine-actionable citations. To address both issues we build upon the foundations of FRBR and CIDOC CRM.

3.1 Managing Primary Texts: FRBR and Canonical Text Services

While we may discover previously unknown documents on inscriptions or in archives, primary sources are by definition finite in number. Every historical document – whether a canonical work or a private contract preserved on papyrus – may appear in multiple editions, multiple translations, and as the subject of commentaries that annotate well-defined extracts of the document as a whole.

Serious users of historical documents need two classes of service. First, they need structured reports about multiple editions and/or translations of, commentaries on, and secondary texts about their historical sources – not just the major canonical works such as the *Aeneid* but every historical document in the public

record. Second, they need better tools with which to understand how these versions of a single work are different. Print editions, for example, only occasionally note where they differ from a previous edition. In a digital library, we can and should be able to see how each edition compares to those previously published and to visualize the relationships between these editions over time. We should also be able to identify translations of any given document, even when these are embedded in unusual places (e.g., translations of short poems from the *Greek Anthology* in a 19th-century magazine). We should be able to compare translations with each other and with their sources. Parallel text analysis not only can help readers of a particular text but provides the foundation for multilingual services such as cross-langage information retrieval and machine translation.

While we need automated methods to track everything published about the Greek tragedian Sophocles, we can create careful metadata about Sophocles, his surviving plays and the numerous fragments of his lost works. Well-curated data about a finite set of documents becomes, in effect, a classification scheme that can be applied to an open set of editions, translations and commentaries.

In 2006, with the rise of Google Book Search and other large scale projects, we realized that we needed to expand our coverage of Greek and Latin source texts. In particular, we needed better data to manage multiple editions of works that were available not only in our collections but as components of image books published by Google and others. In this ongoing work, we have catalogued roughly 585 primary source texts containing 842 distinct authors and at least 1588 individual works. This collection also contains many reference works, including bibliographies, grammars, histories and lexicons. While we found disparate elements with which to create a catalog for Greek and Latin source texts, none was adequate in itself.

MODS records. As reported in [15], our catalog utilizs MODS records downloaded from the LC web service[18] and bibliographic records found in OCLC's WorldCat (used to create MODS records for works not found within the LC Catalog.) Perhaps the greatest challenge is the fact that the majority of our collection can be described as "container works" according to the FRBRoo: a class that "comprises Individual Works whose essence is the selection and/or arrangement of expressions of other works"[9]. A number of our texts fall into this category, such as the *Greek Anthology* (a five-volume series of Greek epigrams with over 100 different authors) or the *Historicorum Romanorum Reliquiae* (a multi-olume set of fragmentary Roman historians, with "works" as short as a paragraph). A large number of volumes contain multiple works by multiple authors, such as a work with selected poems from Vergil, Ovid, and Catullus, or the Greek military histories of Aeneas Tacticus, Asclepiodotus and Onasander. Similarly, even when many books contain only one author, they often have multiple works by that one author, such as the collected orations of Antiphon or the collected plays of Euripides.

Comprehensive domain-specific bibliographies of Greek and Latin. Classicists have long created exhaustive checklists of classical authors – major dictionaries traditionally provide bibliographies of the editions which they used, outlining

broad surveys of Greek and Latin. The Thesaurus Linguae Graecae, Packard Humanities Institute (PHI) and Stoa Consortium have created comprehensive electronic checklists which provide numerical identifiers for a wider range of authors and works than the LC NAF. Unfortunately, none of these lists uses the standard names for authors or works that are already in the LC NAF. Thus, Cicero may be "Marcus Tullius Cicero" in the domain list rather than "Cicero, Marcus Tullius," and Cicero's letters to Atticus may be "Epistulae ad Atticum" rather than "ad Atticum." We combine the LC NAF uniform name with one or more domain specific identifiers: the PHI lists Cicero as author 474, his "Letters to Atticus" as work 57. Neither the MODS records nor the domain specific lists provide a structure within which we can manage multiple editions, much less translations, commentaries, etc. For this we turn to FRBR.

Functional Requirements for Bibliographic Records (FRBR). FRBR is an important data model without which we cannot accomplish the most basic functions on which our user community depends: we need to be able to identify multiple instantiations of primary texts[12]. We wanted to know precisely how many editions, translations, and commentaries of canonical works such as the *Iliad* are in our collection at any one time. We use the FRBR object hierarchy of work, expression and manifestation to represent the *Iliad* as a general work, its various editions, translations and commentaries (which we treat as subclasses of expression) and the various instantiations of these publications such as page images and uncorrected OCR vs. various XML transcriptions (which we treat as manifestations). Related experiments with FRBR and how it might benefit current cataloging practices and digital libraries have proved informative in our efforts reported here[1, 2].

Canonical Text Services (CTS). FRBR is, however, not sufficient for classical studies. Scholars have developed elaborate citation schemes with unique identifiers for particular chunks of text. Strings such as "Il. 3.44" and "Thuc. 3.21" describe book 3, line 44 of Homer's *Iliad* and book 3, chapter 21 of Thucydides. While the precise wording of these chunks will vary from one edition to another and some editors will occasionally redefine the boundaries of particular chunks, canonical citations generally point to the same text in multiple editions. Our digital infrastructure already depends on this foundational knowledge structure that classicists have inherited from earlier centuries. To this end, the CTS protocol has been developed to support more sophisticated querying, organizing, and referencing of texts. The CTS extends the FRBR hierarchy both upwards and downwards, upwards by "grouping Works under a notional entity called 'TextGroup' " and "downwards, allowing identification and abstraction of citeable chunks of text (Homer, *Iliad* Book 1, Line 123), or ranges of citeable chunks (Hom. Il. 1.123-2.22)"[20]. FRBR's "manifestation" may not be relevant for scholarly citation. Therefore CTS focuses more on the semantics of citation practice traditional in fields like classics or biblical studies. We plan to exploit both FRBRoo and CTS as we continue work on our developing catalog.

Our work has thus been fourfold. First, for each "container work" we have created one XML file, which contains both the bibliographic information for

that manifestation and component records for each individual work contained within that manifestation (which include any relevant work identifiers, links to author's online authority records, language information, translator/editor information, etc.) Second, we are creating expression-level XML records for each of these component works within a manifestation (this work is ongoing as we explore means automating this process). The hierarchical nature of XML is quite useful for this kind of bibliographic entity. Third, we are assigning identifiers from standard canons such as the TLG, PHI, the LC NAF, and other relevant bibliographies in order to support the most granular level of text identification possible, a goal of the FRBR-CIDOC harmonization[8]. Where identifiers are not available, we are exploring means of creating them. Fourth, we encode the citation schemes whereby we can extract canonical chunks of text from online documents.

3.2 Managing References by Semantic Class: CIDOC CRM

We have published results from our own work on semantic classification elsewhere[4, 21] (and, of course, the FRBR/CTS work described above entails classification). In this section, we describe one fundamental classification task: mapping fields from two substantial and somewhat overlapping collections on Greco-Roman art and archaeology. This task documents both the importance of semantic classification and the need for entity identification, the fourth layer of Figure 1. For our interchange format, we have chosen the CIDOC CRM, a network-like data structure initiated by the International Council of Museums[7] and accepted as an official ISO standard in 2006.

The CIDOC CRM has evolved over ten years and provides a blueprint for describing concepts and relationships used in cultural heritage documentation. CIDOC CRM can provide an interlingua with which to connect existing data models as well as provide a foundation for new ones. While CIDOC CRM was developed to represent information about objects – especially those managed by museums – a new version of FRBR, FRBRoo, is being developed as an ontology aligned to the CIDOC CRM[9]. FRBRoo provides the means to express the IFLA FRBR data model with the same mechanisms and notations provided by the CIDOC CRM. From our perspective this is a major advance, providing the first third-party integrated data model for textual and art and archaeological collections in twenty years of collection development.

Integration of different cultural heritage vocabularies and descriptive systems is an ongoing research challenge[22]. The CIDOC CRM and FRBR harmonization – especially when extended with the CTS protocol – will allow collections to integrate complex textual materials with rich metadata about objects. Figure 2 shows how two resources – an ancient text passage and a museum catalog object– can be linked together.

As a test case, Perseus has begun collaborating with the German Arachne, the central object database of the German Archaeological institute[10], to create CIDOC CRM records for our art and archaeological collections (5,900 objects and 36,500 images in Perseus, 100,000 objects and 165,000 images in Arachne).

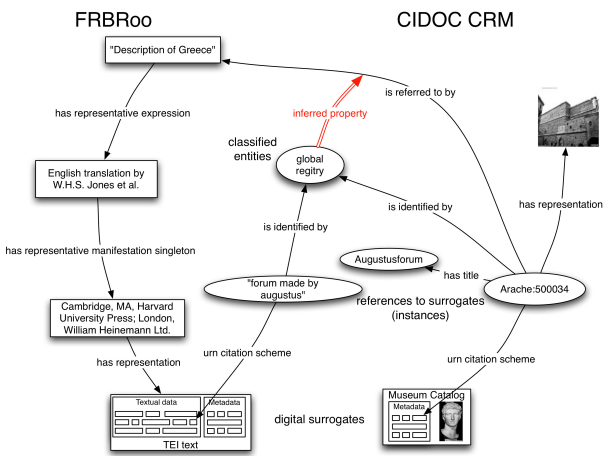

Fig. 2. Entities are linked by using terms of CIDOC CRM and FRBRoo.

This will not only produce a unified database but will link the Greek and Latin collections in Perseus to the same materials in Arachne. Creating CIDOC CRM records from existing metadata is fundamentally a semantic classification task. We map fields labeled x and y in Perseus and a and b in Arachne to m and n in CIDOC CRM. Both Perseus and Archne implement specialized data models that have been refined to meet the needs of a specific perspective. They do not directly conform with standard metadata schemas and therefore have to be manually mapped to a data model that conforms to the classes and properties of the CIDOC CRM[14].

While the complexity of the CIDOC CRM data model poses difficulties for conversion, Figure 3 illustrates the even deeper challenges that we face. In this case where we have two records for the same object, we can see where semantic alignment introduces questions of data analysis and fusion. The problem of language appears at once – we need to establish that "bust" and "Portraitkopf" are English and German equivalents. We also need to address variations where language is not a factor. For example, we need to match "H 44 cm" in Arachne with "H . 0433 m" in Perseus – two comparable but not quite equivalent figures for the height of the bust. Augustus is the same in German and English but none of the data presented unambiguously indicates that this is "Augustus, Emperor of Rome, 63 B.C.-14 A.D." We find variant spellings of the placename Aricia/Ariccia in each record. More significantly, the Perseus record "Aricia, near Rome" provides a clue that a named entity system could use to establish that this Aricia corresponds to tgn,7007011[1] in the Getty Thesaurus of Geographic Names[13]. We want to be able to recognize that "Boschung 1993" in Perseus and "D . Boschung, Die Bildnisse des Augustus" refer to the same bibliographic entity.

[1] "Ariccia [12.683,41.717] (inhabited place), Roma, Lazio, Italia, Europe."

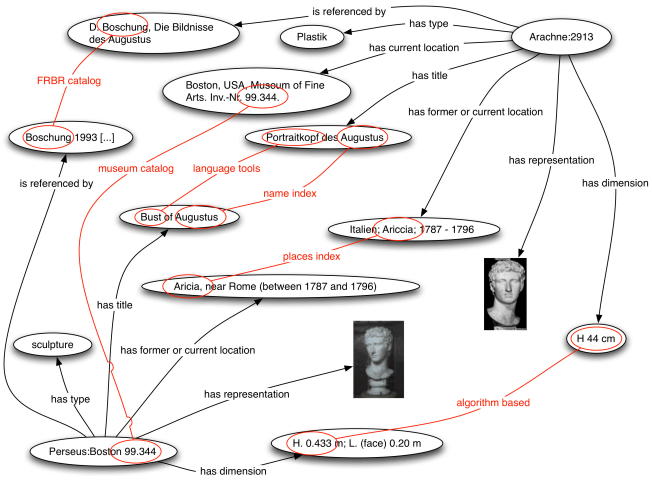

Fig. 3. Identification of instances that refer to the same entity.

Semantic classification thus raises the problem of entity identification. Knowing Alexandria is a place is useful, but not nearly as useful as knowing that a particular Alexandria refers to the famous Egyptian city (tgn,7001188).

4 Named Entity Identification

In twenty years of digital collection development, we have consistently stressed developing knowledge sources that not only enumerate but provide machine-actionable descriptions of domain-specific entities. Many comprehensive knowledge sources exist in print form that can be converted into machine-actionable resources to provide new services for users. Our own group has focused on creating such machine-actionable knowledge sources and accompanying services[5, 6]. As we move towards a VRE for Greco-Roman studies, we have focused on three sources.

1. *Markup of primary sources:* Multiple editions of the same work can be aligned against one another, collated and used to correct each other. Careful markup from one edition of an author, however, can be projected onto other editions as well. Thus, while the number of editions for any historical work may be an open set, a single well-structured version can become a template to structure many other editions. In a field such as classics, primary source citations are probably the most important single entity – if we can recognize and decode strings such as "Thuc. 1.38" as references to Thucydides' book one, chapter thirty eight, we can generate links between many different works and analyze scholarly trends. If we have one full text marked up, we can then often identify floating quotations even when there are no recognizable citations nearby.

2. *Author indices:* Multiple indices exist for most classical authors – thousands of indices with hundreds of thousands of entries and millions of references. These indices store the judgments of experts as to which Alexander or which Alexandria is meant in a given passage. A baseline name identification strategy based on document indices produces good results because most names are unambiguous in a given document – real ambiguity occurs as documents grow in size or are combined. If we simply connect every ambiguous entity to the most common entity with that name, we are successful 97.4%, 95.3%, 93.5% and 91.7% of the time for Thucydides, Herodotus, Pausanias and Apollodorus, respectively. Thucydides has the most well-defined subject and it is thus not surprising that this baseline method performs best with his work and worst with the much more heterogeneous Apollodorus. However, if we combine Herodotus, Pausanias and Apollodorus into one large document, the overall accuracy of this baseline technique for those authors falls to 91.4%. This reinforces the intuitive assumption that this baseline method becomes less accurate as documents increase in size (thus increasing the probability that ambiguous names will appear).
3. *Reference works:* In fields where canonical citation schemes map source texts, we find not only the usual textual descriptions of people and places, but citations that associate passages of particular texts with the current article. Such encyclopedias thus constitute broad indices of major entities across a specific domain. We concentrated on two reference works: Smith's *Dictionary of Greek and Roman Geography* for place names (which contains 11,564 entries and has yielded 25,748 citations) and Smith's *Dictionary of Greek and Roman Biography and Mythology* for personal names (which contains 20,336 entries and has yielded 37,549 citations). Together, they provide a broad framework for the field.

In the 300-volume Perseus American collection, roughly 25% of the books have indices – as very large collections include millions of books we will need to consider how best to mine the information from millions of indices as well as many thousands of reference materials[3]. In a field such as classics, however, the canonical citation schemes available for most authors provides citations that are not only useful in themselves but that contribute to a major development challenge: we need to integrate the deep information available in individual author indices with broad resources such as the Smith's dictionaries.

Figure 4 shows a personal name entry in the Perseus Encyclopedia (PE), that for Abderus, a son of Hermes, identified by "abderus-1,' while Figure 5 shows a similar entry from the Smith *Dictionary of Greek and Roman Biography and Mythology*, where the identifier for the same Abderus is "abderus-bio-1." Using the context of both entries, our automatic system was able to correctly map the entries from these two resources to each other, identifying PE "abderus-1" as the Smith "abderus-bio-1." While this example is far more straightforward than most, it serves to illustrate the type of matching being performed.

The system achieved an overall accuracy of 78.6% when aligning these two resources. Table 1 provides an overview of the tagging accuracy of the system

```
<div1 type="entry" id="abderus">
  <head>Abderus</head>
  <div2 type="subentry" id="abderus-1">
    <head><persName>
    <surname>Abderus</surname> </persName>,son of <persName><surname>Hermes</surname>
    </persName></head>
      <div3 type="index"><list type="index">
       <item>killed by the mares of Diomedes: <bibl n="Apollod. 2.5.7">Apollod.2.5.7</bibl></item>
       <item>the city of Abdera founded by Herakles beside his grave: <bibl n="Apollod.
              2.5.7">Apollod. 2.5.7</bibl></item>
          </list></div3></div2></div1>
```

Fig. 4. Personal Name Entry in PE XML file

```
<div2 type="entry" id="abderus-bio-1" org="uniform" sample="complete">
    <head><label>ABDE´RUS</label></head>
    <p>(<label lang="greek">*)/Abdhros</label> ), a son of Hermes, or according to others of
       Thromius the Locrian. (<bibl n="Apollod. 2.5.8" default="NO" valid="yes">Apollod.
       2.5.8</bibl>; Strab. 7. p. 331.) He was a favourite of Heracles, and was torn to pieces by
       the mares of Diomedes, which Heracles had given him to pursue the Bistones. Heracles is said
       to have built the town of Abdera to honour him. According to Hyginus, (<bibl n="Hyg. Fab. 30" default="NO"
       valid="yes">Hyg. Fab. 30 </bibl>,) Abderus was a servant of Diomedes, the king of the Thracian Bistones, and was killed by
       Heracles together with his master and his four men-devouring horses. (Compare Philostrat. <hi rend="ital">Heroic.</hi>
       3. &sect; 1; 19. &sect; 2.)</p>
    <byline>[<ref target="author.L.S" targOrder="U">L.S</ref>]</byline></div2>
```

Fig. 5. Personal Name Entry in Smith

across all entity types. The system achieved similar results for personal and place names with significantly lower performance for ethnic groups. When the system correctly found no match it meant that the PE entity had no relevant match in Smith. The category of errors reflects when either an error in the PE or in the Smith XML file caused a tagging or other type of error. Occasionally, a match had to be marked as uncertain, due to insufficient textual content. It required 1,000 hours of labor to align 9,000 entities but the resulting unified database of disambiguated reference to entities in texts is c. 100,000. We should emphasize that these 100,000 disproportionately identify less common entities: our indices contain a far larger percentage of references to the lesser Alexandrias than to the famous city of that name in Egypt. The bias of these 100,000 entries provides broader coverage than a random sampling of 100,000 entries would contain, since a random sampling would contain more references to very common (and less frequently indexed) names such as Alexandria. In real work, users need help finding these obscure entities - they want to find references to one of the smaller cities that Alexander founded and named after himself. If 95% of our Alexandria references point to Egypt, digital libraries only begin to add value insofar as they help us locate the ten ancient Alexandrias that make up most of the remaining 5%.

5 Conclusion

Recent steps towards a VRE for Greco-Roman antiquity strengthens our long-term belief that any effective digital infrastructure must address the entity problem at various levels. First, library catalogue records in classics not only need to include more authors and works but they need to incorporate canonical citation

Table 1. Alignment Accuracy by Entity Type

Category	**Total**	**Ethnic**	**Place**	**Personal**	**Other**
Corr. found no match	3065	111	356	1421	1178
Error	14	1	10	3	0
Uncertain	5	0	5	0	0
Incor. matched	1880	217	423	1156	83
Corr. matched	3950	65	1221	2660	4
Tot. entities	8914	394	2015	5240	1265
Accuracy	**78.69%**	44.89%	78.26%	77.86%	93.44%

schemes – new classes of entity – if they are to provide the foundations for serious digital libraries. In the first generation of digital collections, classicists for the most part ignored catalogue records as being incomplete and, from their perspective, static. Second, using the CIDOC CRM to unify large collections of data provides an important and useful first step but immediately raises the problem of entity identification: we need to be able not only to recognize that English Athens and German Athen are equivalent but to distinguish Athens, Greece, from Athens, Georgia. Third, document indices, encyclopedias, gazetteers and other reference tools contain vast amounts of named entity identification data of the form "entity-X occurs at location-Y." These sources provide information of immediate value to human readers and potential training data for machine learning. Improved tools with which to merge this data should be a major priority of cyberinfastructure. While these conclusions reflect work on a particular domain within the humanities and stress textual materials, all intellectual discourse bases its arguments on meaningful entities extracted from raw data. We need to move towards a generalized architecture that supports named entity services for engineering and the social and natural sciences as well as the humanities.

References

1. T. Aalberg, F. B. Haugen, and O. Husby. A tool for converting from MARC to FRBR. In *ECDL*, v. 4172 of *Lecture Notes in Computer Science*, pp. 453–456. Springer, 2006.
2. G. Buchanan. FRBR: enriching and integrating digital libraries. In *JCDL '06: Proc. of the 6th ACM/IEEE-CS Joint Conf. on DLs*, pp. 260–269. ACM Press, 2006.
3. G. Crane and A. Jones. Perseus American Collection 1.0. Tufts DL, 2005. `http://dl.tufts.edu//view_pdf.jsp?urn=tufts:facpubs:gcrane-2006.00001`.
4. G. Crane and A. Jones. The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In *JCDL '06: Proc. of the 6th ACM/IEEE-CS Joint Conf. on DLs*, pp. 31–40. ACM Press, 2006.
5. G. Crane and A. Jones. Text, information, knowledge and the evolving record of humanity. *D-Lib Magazine*, 12(3), 2006.
6. G. Crane, et. al. Towards a cultural heritage digital library. In *JCDL' 03: Proc. of the 3rd ACM/IEEE-CS Joint Conf. on DLs*, pp. 75–86, Houston, TX.

7. N. Crofts, et. al. Definition of the CIDOC object-oriented conceptual reference model. Technical report. http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.pdf.
8. M. Doerr and P. Le Bouef. Modelling intellectual processes: the FRBR-CRM harmonization. In *DELOS Conf. on DLs*, Tirennia, Pisa, Italy, 02/2007.
9. M. Doerr and P. LeBouef. FRBR: Object-Oriented definition and mapping to the FRBR-ER. 02/2007. http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR_oo_V0.7.0.pdf.
10. R. Förtsch. ARACHNE - Datenbank und kulturelle Archive des Forschungsarchivs für Antike Plastik Köln und des Deutschen Archäologischen Instituts, 2007. http://arachne.uni-koeln.de/inhalt_text.html.
11. P. Gietz, et. al. TextGrid and eHumanities. In *E-SCIENCE '06: Proc. of the Second IEEE International Conf. on e-Science and Grid Computing*, pp. 133–141, Wash. DC, USA, 2006. IEEE.
12. IFLA. *Functional Requirements for Bibliographic Records: Final Report*, volume 19 of *UBCIM Publications-New Series*. K.G.Saur, München, 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf.
13. Getty TGN. http://www.getty.edu/research/tools/vocabulary/tgn/.
14. R. Kummer. Integrating data from the Perseus Project and Arachne using the CIDOC CRM: An examination from a software developer's perspective. In *Exploring the Limits of Global Models for Integration and Use of Historical and Scientific Information-ICS Forth Workshop*, Heraklion, Crete, 10/2006. http://www.perseus.tufts.edu/~rokummer/KummerCIDOC2006.pdf.
15. D. Mimno, G. Crane, and A. Jones. Hierarchical catalog records: Implementing a FRBR catalog. *D-Lib Magazine*, 11(10), 2005.
16. Library of Congress. MADS. http://www.loc.gov/standards/mads/.
17. Library of Congress MODS. http://www.loc.gov/standards/mods/.
18. Library of Congress. http://z3950.loc.gov:7090/voyager?.
19. ACLS Commission on Cyberinfrastructure. Our cultural commonwealth: The final report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences, 2006. http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf.
20. D. Porter, et. al. Creating CTS collections. In *Digital Humanities*, pp. 269–74, 2006.
21. D. A. Smith. Detecting events with date and place information in unstructured text. In *JCDL '02: Proc. of the 2nd ACM/IEEE-CS Joint Conf. on DLs*, pp. 191–196, New York, NY, USA, 2002. ACM Press.
22. M. van Gendt, et. al. Semantic Web techniques for multiple views on heterogeneous collections: A case study. In *ECDL*, volume 4172 of *Lecture Notes in Computer Science*, pp. 426–437. Springer, 2006.
23. G. Weaver, B. Strickland, and G. Crane. Quantifying the accuracy of relational statements in Wikipedia: a methodology. In *JCDL '06: Proc. of the 6th ACM/IEEE-CS Joint Conf. on DLs*, page 358. ACM Press, 2006.