1. Automatic Metadata Generation and Evaluation

Good “quality metadata supports the functional requirements of the system it is designed to support” (Guy, Powell, & Day, 2004). In the case of DRIADE, metadata is essential for preservation, resource discovery, and eventually data integration—particularly as the initiative grows. Metadata generation and quality evaluation are required to not only develop but to maintain and carry on DRIADE. A research effort in this includes identifying, evaluating, and implementing both automatic metadata generation and automatic quality evaluation capabilities to address metadata challenges that will threaten to stem DRIADE’s growth and health.

Accurate, coherent, reliable metadata is required if DRIADE users are to find and also use data objects (see Burce & Hillman, 2004 for more on metadata criteria). Metadata quality will be a growing challenge for DRIADE due to the different participants generating metadata (scientist/data creators/authors) and to the overall effectiveness of any automatic algorithms/s we implement. It’s possible, although less likely given DRIADE’s plans, that the volatile nature of data objects in digital form will also impact metadata quality. Regardless, inaccurate and obsolete metadata will have a negative impact on DRIADE’s effectiveness.

DRIADE will present a challenging, and to some degree, a changing environment, demanding high quality metadata, but it will likely be too difficult to rely on humans alone to curate and evaluate metadata. In conducting research in this area we should look to develop a sound metadata generation process model integrating human and machine capabilities. Research in this area will help us to determine:

1. How much we can rely on automatic capabilities—and satisfy our goal to use machine processing as much as possible for metadata generation.
2. How to effectively use automatic processes to prompt and guide scientists when they need to create metadata.
3. How we can limit curators’ (metadata professionals) tasks to metadata generation/evaluation activities that truly require their expertise.

There is absolutely no reason a data curator should be typing in the name of a person who authored an article or created a data set, when this metadata can be automatically extracted from a resource via semi-structured metadata processing, inherited from another metadata record, or possibly harvested from the data object. In the area of automatic metadata generation, we can look at methods of extraction, harvesting, and derivation based on system protocols.

- **Extraction** employs automatic indexing techniques to resource content to generate metadata.
- **Harvesting** automatically collects metadata previously tagged by humans or automatic processes.
- **Derivation** automatically creates metadata according to a stored system profile.

(Greenberg, et al, 2006)
The AMeGA study’s Recommended Functionalities for Automatic Metadata Generation Applications, Version, 1.0 may be good place to start (http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf). In the area of automatic metadata quality evaluation, we can look specifically at section 7 of the AMeGA report (beginning on page 53) to begin to determine means for which we can create a statistical rating score. A confidence interval could be used to determine the overall quality of the metadata, and if or which metadata might require human evaluation. This process will require a great deal of testing, but I believe it is very important to pursue. Consider Lesk’s statement about the cost of metadata. Automatic metadata generation and evaluation will (eventually) reduce these costs. Note that automatic metadata generation for data objects, at the level that will help DRIADE, has not been a primary research area; we are in a unique position with DRIADE to take advantage of the close coupling of published research and underlying data objects to study this problem. Even so, we need to be mindful of the limitations of this approach.


2. Dynamic Vocabulary Integration & Maintenance

Digital technologies and networks have revolutionized the development of information services, and the Web is now the home to numerous digital library, cultural resource/archival collections, and increasingly data repositories—many of which are interdisciplinary. DRIADE fits this paradigm, as evolutionary biology is an interdisciplinary field that will be Web accessible.

The rise of interdisciplinarity and incorporation of new technologies challenge traditional research approaches, sometimes making the research task more difficult, particularly in environments where resource collections cover multiple disciplines (National Research Council, 2001). Pellmar and Eisenberg (2000, p. 51) attest that “language jargon barriers…continue to haunt those developing search engines for online publications.” With DRIADE, we can think of access challenges presented by the diversity of vocabularies supporting the various disciplines comprising evolutionary biology, as well as the range of researchers we might support (the armature researcher, the doctoral student, and the seasoned scientists). Digital technologies can promote multidisciplinary and interdisciplinary research, but this can only be successful if the information infrastructure provides access across domains—including efficient and effective means for integrating vocabularies to facilitate indexing and access.

* Grid computing research has made progress here, but I need to look further at what specifically has been accomplished and see what fits with DRIADE.
Research in the areas of vocabulary/ontological integration could include:
1. Studying the means by which people are gaining access (and not) to research and data for interdisciplinary research/evolutionary biology.
2. Evaluating approaches to vocabulary (and perhaps ontological) integration.
3. Developing and demonstrating an approach for dynamically integrating existing domain-specific vocabularies using Semantic Web technologies. The goal here would be to develop a dynamic vocabulary server.
4. Examining the impact of the integrated ontology on retrieval effectiveness—measuring precision, recall, user satisfaction. This could also include examining the use of the integrated vocabulary for query expansion.

A related area that we *might* explore would be the development of vocabulary registries, and the curation and maintenance of concepts.


3. Instantiation/Data Object Families

Scientists create data objects through a variety of methods (observation, experimentation, interviewing), using a variety of software and means for capturing data. At times they copy/reproduce and store an exact copy of one data object in another system or use another software package (e.g., a data set may be created in Excel, but copied over to SPSS). Without any change, the instantiations might be explained as equivalent sets. Scientists also revise, merge, synthesize, and perform other on data object to create a new object. A scientist may use a subset of one data object as a basis for creating another objects. If data object A serves as a primary foundation for producing data object B, and elements of A are very obvious, B might be considered a derivative of A. As part of DRIADE’s metadata research plan, we might ask to what extent is it important for a system like DRIADE to track and make explicit the relationship among data objects over a life time? We will be exploring this in the DRIADE study we are planning. (See DRIADE research plan document for ideas on this, including diagrams)