

## Briefing Paper: developing the DOI Namespace

*This briefing paper describes a project that has been commissioned by the IDF for completion during the first half of 2001. The paper describes the scope of the project, and its anticipated outcome. It is a public paper, intended for circulation to any interested party.*

### Background

The IDF has recognised the need for the development of **Application Profiles** (DOI-APs) for the management of IP entities identified with DOIs. A significant element of DOI-APs will be metadata schemas, designed to support a specific application or set of applications that are regarded as having value by the DOI User Community responsible for their development. This project is concerned only with building a necessary toolbox for the development of these metadata schemas, and not with other aspects of the development of DOI-APs.<sup>1</sup>

The IDF accepts the <indec> analysis for the development of its metadata requirements, in particular the need to ensure long-term interoperability between different schemas. The <indec> analysis provides the necessary intellectual framework for the development of such interoperability, but practical implementation was not part of the project. If the IDF wishes to use this framework to ensure interoperability between schemas, then it has to develop its own implementation. This document describes the project that has been commissioned for the initial development of the “**DOI Namespace**” (DOI-NS).

It is important to recognise that initial development is only the first step.

### Tasks involved in developing a metadata registry – overview

The implementation of an <indec>-based metadata registry as a tool for metadata schema development implies that the following steps will need to be taken:

- The development of a data dictionary (DOI-NS), structured on <indec> principles,<sup>2</sup> with controlled values for all significant attribute types that we anticipate are likely to be needed in DOI-AP metadata schemas in the near future.
- The publication of this data dictionary in a format that allows developers access to it.
- The development of a process for continuing maintenance and development.

We anticipate that the development of a “metadata registry” for automated online interoperability is likely to be the ultimate outcome of the development of a DOI-NS. However, such a development is certainly not within the scope of this project, which is concerned only with the compilation of the data dictionary and its publication in a human accessible form. Nevertheless, we will do what we can<sup>3</sup> to ensure that the DOI-NS as developed under the terms of this project plan does not in any way conflict with its future deployment within an automated metadata registry implementation.

---

<sup>1</sup> For a complete description of the elements necessary for the complete definition of a DOI-AP, please see the DOI Handbook V1.0 (in preparation).

<sup>2</sup> Structuring on <indec> principles has a number of implications, designed to ensure unlimited extensibility. One of these is that only a single attribute should change at each level of the hierarchy. Secondly, terms exist not in a single structured hierarchy but in a genealogy. Whether it is necessary to develop genealogical definitions based on the <indec> genealogical syntax is a matter for discussion. This syntax is likely to have value only in the construction of currently somewhat notional automated interoperation applications; it would be theoretically possible to retrofit such definition structures, but ultimately this would be much more time consuming. The use of genealogical structuring ensures that terms are being added within the logical framework but inevitably increases the time that the dictionary development will take.

<sup>3</sup> Including, for example, compliance with ISO 11179 assuming that this can be achieved without compromising our other objectives.

## Developing a data dictionary

It is not necessary to develop the data dictionary from scratch. <indec> itself has a published data dictionary of over 200 terms, and a further unpublished dictionary exists of nearly 1000 terms to which we have access. The EPICS data dictionary includes a few hundred higher-level terms, and a substantially greater number of terms in value lists. Furthermore, a number of appropriate value lists already exist, managed by outside authorities, including for example ISO.

However, significant tasks remain:

- The unpublished parts of the <indec> data dictionary are known to have been developed in a somewhat random process and have never been systematically reviewed or tested. Furthermore, the presentation of this data dictionary (which is constructed hierarchically across many columns of an Excel spreadsheet) makes it extremely difficult to use and comprehend. **The task** is to extract the content of this document, to allow a full review of the extended <indec> data dictionary, extracting from it the terms that are relevant to the DOI-NS, and validating the existing structure.
- We know that there are certain differences between <indec> and EPICS/ONIX that must be resolved if the two data dictionaries are to be satisfactorily used together. Some of these differences relate to the particular use of language; while these may be time-consuming to resolve, they are not likely to be major obstructions to reaching a final solution (they can often be resolved by coining neologisms or even using meaningless placeholders in the short term). There *may* however be more fundamental problems. In either case, the IDF is developing a review process for resolving such difficulties. **The task** is to bring relevant terms in EPICS/ONIX into a common DOI-NS with the relevant terms from the <indec> data dictionary, resolving any ambiguity and disagreement between the two data dictionaries in the process.

Once these two tasks have been completed satisfactorily, the gaps that need to be filled must be identified. We have already identified<sup>4</sup> that controlled vocabularies are required for DOI-AP metadata schemas in:

- Agent role codes
- Quantities and measures
- Component and derivative codes
- Rights ownership and management codes (this is a very small subset of what will be required for the development of a full rights metadata infrastructure, and will have to be somewhat provisional if developed separately)

As part of this process, we will also have to manage a process of unique identification of the terms in this dictionary with an identifier (“token”), which is linked to the definition rather than to the headword.<sup>5</sup>

We cannot at this time predict with great precision the total number of terms that will be included in this initial development phase. Our aim will be to ensure that we have covered the requirements of what we can currently predict are likely to be the needs of DOI-APs in the relatively near future, and built a structure into which new requirements can be incorporated

---

<sup>4</sup> See Appendix 3 of the DOI Handbook V1.0 (in preparation).

<sup>5</sup> Ultimately, it is the definition that is important rather than the headword, which is why the definition is identified rather than the headword itself. Of course, the headword once used becomes an identifier in its own right and like any other identifier should only be reused with a different meaning in wholly exceptional circumstances. We have decided (subject to final confirmation) to develop a set of DOI-NS identifiers rather than to attempt to adopt existing identifiers like the iid (indec item identifier). This has been driven by simplicity of process rather than being a decision taken on grounds of principle.

relatively straightforwardly. This includes creating all controlled vocabularies listed above, although these may not be comprehensive for all media types.

The work will draw not only on <indecs> (unpublished as well as published) and EPICS, but also on the extensive work on images as part of Steffen Lindek's work for Bioimage, and on the CrossRef metadata set. One key task will be to ensure the delivery of a complete set of values for use within each of the elements of the DOI Kernel (including solving the vexed question of "digital" and "physical" manifestations, which we believe to be a matter simply of semantics rather than any fundamental disagreement over the requirement).

We also believe that there will be a requirement for the development, naming and unique identification of complex structures ("composites") for the standard structuring of attribute groups where these structures are likely to be needed in several metadata schemas. EPICS/ONIX already encompasses a significant number of composite structures that are now being implemented in an international environment and we would use these so far as possible as the basis for common IDF data structures.

There are significant complexities in performing the task that we should acknowledge here and that makes some aspects difficult to quantify fully. The <indecs> data dictionary was built from top down, whereas a number of aspects of the EPICS data dictionary, particularly some of the code lists, were built from the bottom up. We cannot at this stage do anything other than guess at the difficulties that will be encountered in attempting to cross the middle ground – it may turn out to be relatively trivial or it may turn out to be extremely difficult. Our best guess is that the answer will turn out to be "both" – some aspects of the dictionary will be easy to complete, some very difficult indeed.

## **Developing a technical infrastructure for managing the data dictionary**

The use of a spreadsheet for managing a data dictionary of this kind is self-evidently unsatisfactory, and there is no question that some sort of database needs to be developed. This database needs to be a production tool, but will also need to double as a publication tool. As such, the interface to the database needs to be such as to allow simple navigation through the structure and to ensure that the structure of the dictionary can be displayed graphically as well as lexically.

An early part of the project involves a review of pre-existing tools, to discover whether there are any that will meet our (fairly simple) requirements.

**The task** here can be broken down into three phases:

1. Developing a functional spec for the database. This task has already (in January 2001) been started. A first pass has been made at defining the data that needs to be captured relating to each defined vocabulary term in the database, and the necessary structures to support linkage between items in the database. An initial pass has also been made at identifying the data input, maintenance, access and output functions that need to be supported.
2. Converting the functional spec into a full system spec. This process, which is just beginning, involves considering what packaged tools may already be available that might assist with what is essentially a lexicographical process. It will also consider whether the production database and the publication database were essentially "the same thing" or whether they should be differently structured. It will also involve selecting an appropriately qualified candidate to perform the technical work of database construction.
3. The final aspect of the technical work will involve building and deploying the system itself.

## Developing a continuing maintenance and development process

A core element of the project will be to recommend a process for continuing maintenance and development of the content of the database, as well as its technical management. This could involve continued outsourcing, or the development of an in-house metadata function within the IDF. **The task** will be the delivery of a report, with recommendations.

### Timetable

1. **Ordering of tasks:** the first task is to complete the functional spec for the data and the database; this has to be completed before any substantive work can begin on the data preparation tasks, or on the technology tasks. The data preparation and technology tasks can then continue in parallel, although clearly completing the data dictionary becomes dependent on the completion of at least the production aspects (data input and maintenance) of the database. The report with recommendations for the ongoing development and maintenance of the database will be delivered at least 2 months prior to the completion of the project.
2. **Milestones – technical development:** We can identify certain initial milestones for the project; we expect to be able to complete the requirements spec before the end of January; the development of a prototype database, with the necessary functionality for data entry and review, does not need to be complete before work on the data dictionary begins, but its development may become a rate limiting step. The work on this aspect of the project is therefore very urgent.
3. **Milestones – data dictionary development**
  - a. **Part I:** <indec> and EPICS/ONIX. The first step will be to review the basic <indec> vocabulary and to identify those aspects that require improvement or extension. This involves name space (NS) issues as well as controlled vocabulary (CV) issues. The second step will be to review the EPICS/ONIX vocabulary, to identify those parts that are relevant for the DOI work and to map them onto the <indec> schema. This includes the generation of a genealogy and a revision of the terminology (EPICS/ONIX vs. the basic <indec> vocabulary). The third step will be to review the extended <indec> vocabulary and to look for terms that can be used to incorporate EPICS/ONIX metadata into the <indec> framework.
  - b. **Part II:** The second part will consist of a review of those schemas that have already been developed for DOI-APs: CrossRef and Bioimage. These will be used to extend and improve the DOI data dictionary (both for NS and CV issues).
  - c. **Part III:** The third part is related to the identification and creation of CVs and special elements such as composites required for DOI work insofar as they have not been subject to development in Parts I or II. In this phase, the schemas are completed as much as possible.
  - d. We do not expect that these tasks can be dealt with in a strict sequence; the development of the data dictionary will involve iterative refinement (either cycling between Parts I, II and III, or cycling within these parts). We expect the duration of this process for the primary developer to last approximately 12 weeks.
  - e. Allowing for a maximum of three weeks of work on dictionary development in each elapsed month (leaving time for intermediate reviews) the entire project should therefore be finished by the end of May.

## Methodology and resources

1. The project is being undertaken by a group of consultants with existing relationships to the IDF, who have been close to many of the developments on which the DOI-NS will be built.
2. The development of the technical infrastructure to support the DOI-NS is being managed by Eamonn Neylon, who is also the overall project manager.
3. A single individual, supported by a team of reviewers, can best manage the primary **data dictionary development** and production tasks. Steffen Lindek, who has extensive experience of developing metadata schemas with the Bioimage project, is taking this primary role. David Martin (ONIX/EPICS) and Mark Bide (<indec>) are providing the first source of advice and support. The IDF is establishing a metadata review panel, and is seeking additional members to join Steffen Lindek, David Martin, Mark Bide, Eamonn Neylon and Norman Paskin. This panel will provide the decision-making forum for the resolution of issues where the primary group is unable to reach consensus, or finds a need for wider input to a particular problem. Although the group will primarily be an email discussion group, we believe that it is likely to need to meet in person at least a couple of times during the lifetime of the project. Each member of the review panel should expect to devote at least one day each month to the project.
4. EN, MB, SL and DM will provide a **report with recommendations for ongoing data dictionary development and maintenance** to be presented to the review panel in the third month of the project.