nature webdebates

A view from the news industry

David Allen
Managing Director, International Press Telecommunications Council

Papers in scientific journals and news items at first sight may not have much in common. But many of the technical challenges now being faced by scientists have already been the focus of intense debate in the news industry where consensus solutions are now emerging. The Internet means that although resources are located all around the globe this is no longer a barrier to seamless access and retrieval. The latter requires that data must be identified and managed in a consistent manner, and it is significant that the highly competitive press has managed to agree on consensual standards across the industry, recognising as it has that such standards will improve functionality and reader-experience for all newspapers and the industry as a whole. The long-term preservation of both important scientific and news material is also best ensured through distributed storage strategies, as opposed to centralized repositories.

Historically, scientific information has been stored in many formats. In some cases it is only available on paper or film. Similar problems are faced in the news industry and we have been working on the development of standards to help improve information exchange in the digital era. News can originate from anywhere and may be presented for distribution in various formats and languages, and the size of the scientific literature pales beside the enormous volumes of material that need to be handled in the news industry. We needed to consider how to deal with such a variety of source material in a consistent and efficient manner. Similar standards can also be used for the standardised retrieval and exchange of scientific papers and other data even if the repositories are widely separated. But achieving this goal will require the scientific community and others involved in the dissemination of research results following the news industry's lead in agreeing to consensus standards.

We chose to base our work on eXtensible Markup Language (XML), an open language for representing machine-readable data recommended by the World Wide Web Consortium. Its key features are that it:

1. is an open standard, and not proprietary to this or that publisher
2. can be manipulated using open-source tools
3. allows documents to be tagged with rich metadata (information about the document and its contents), and can be easily customised.
4. can be used to describe almost any sort of content, from text to video
5. is both human and machine readable.

At present the bulk of news and scientific literature is in text form. Our first standard was the News Industry Text Format, which was designed specifically to address the issue of marking up news stories. It has a simplified structural model consisting of a head and body with the content broken into blocks that may contain paragraphs or tables. However, it has also a rich set of content mark-up to identify specific people, places, dates, organisations and other key entities in the news domain. In contrast, the scientific world is only waking up to the fact that such tagging of documents is key to improving search, retrieval, and matching across disparate sources.

More recently, the advent of widespread multimedia and recognition of the need to cope better with the dynamic and evolutionary nature of news prompted us to look at a more comprehensive model, designed to cope with arbitrary media and with multiple objects. This resulted in the NewsML standard that was published in late 2000. NewsML has a rich set of metadata. It can be used to describe relationships between different items of content, is highly customisable and is independent of the content media type.

This modern approach to news necessitates a rich set of metadata to be used during the content lifetime. We have broken the NewsML metadata into three specific types:

1. Factual data that gives concrete information about the content in terms of its origin, format etc. This is termed AdministrativeMetadata.
2. Legal conditions covering copyright, distribution and publishing rights and is called RightsMetadata.
3. Contextual inferences that may describe what the object is about, who it is intended to be targeted at, etc. This is named DescriptiveMetadata.

There is a fourth, less specific, type that is a general extension metadata. This may be employed by individuals press organisations to define any additional categories of metadata that they may require to meet their specific needs.

Within NewsML it is possible to describe relationships between multiple objects of very different forms of media. These relationships can be described using specific predefined terminology. Where objects are presented in different forms (for example, in language or image resolution) a basis for choice may be provided for the final user.

Customisation is an important feature and may be achieved by specifying and publishing controlled vocabularies relevant to the user domain of the data. Default vocabularies have been defined for the news domain; controlled vocabularies have been developed in science for library indexing, but there is a need to develop new ones more tuned to the needs of content management on the Web. It is encouraging to see renewed interest in standards, controlled vocabularies, and ontologies, both in science, and in the computing science community.

The design aim was that the content object should be described by its metadata once and that this metadata could be transformed for use within databases or other repositories, by press groups with specific needs. Physically based source material will need to be converted as necessary to make it suitable for scanning and rendering into PDF or other display formats. These can be read subsequently without the need for the scanned data to be converted into character-based strings. This approach enables a uniform way of managing, accessing, identifying and exchanging information.

Further information on these standards may be found at the IPTC web site.