Who said that?

Gordon Dunsire

Presented at 13. seminar Arhivi, Knjižnice, Muzeji

25-27 Nov 2009, Rovinj, Croatia

Published by Gordon Dunsire

Edinburgh 2012

Who said that?

Presented at 13. seminar Arhivi, Knjižnice, Muzeji
25-27 Nov 2009, Rovinj, Croatia
Gordon Dunsire

Metadata

Metadata are statements about information, especially when the information is packaged. The package is often referred to as an information resource; packaging containers take many forms, including printed books, websites, digital versatile disks, etc. Metadata is therefore data about data; "a library catalog is metadata because it describes publications" [1]. Metadata itself is information and it is possible to make statements about it: metadata about metadata.

The statements which constitute metadata can be made in regular, structured patterns, but this is not an intrinsic constraint and metadata can take the form of single statements describing multiple aspects of the resource. A common example of unstructured metadata is the abstract describing a scholarly work. Professional information communities prefer to take a structured approach, in the form of archival descriptions, library catalogues, and museum documentation, to make it easier for their clients to use this metadata and to support its creation, maintenance, and distribution. A separate but closely connected issue is the use of language in metadata statements. Again, this can vary from tight control, for example using authority systems, to few or no constraints, such as in a back-of-the-book index.

Who makes such statements? The simple answer is anyone who has the capability of doing so, although an interest in the information resource is an added incentive. Before the World Wide Web, such capability was confined largely to information professionals working in archives, libraries, museums, and publishing. Characteristics of this activity include a high level of training and skill applied in a general context and objective fashion. It should be noted, however, that abstracts, topic keywords and similar metadata are often created by authors. The ubiquity of the Web has given information consumers much greater opportunity to create metadata, manifested in social networking or Web2.0 services. We can refer to members of the public, our end-users and audience, as information amateurs. They are relatively untrained and lack the skills required to produce metadata effective over a wide range of resources for a broad range of consumers; their metadata are often intended for a narrow, personal context, but there is nothing wrong with that. Continuing with this metaphor, we can identify a third category of metadata creators, the information stupids. These are not people, but computers, and significant quantities of metadata are now produced by machine. A full-text index of an electronic document is metadata, whether permanent or created on the fly,: at the very least, such an index indicates that a specific word can be found in the document, and constitutes a statement about the information contained in the document. It is the basis of Google and many other Web search engines. Computers are really stupid because they have no brain or even something resembling a human brain ("electronic brain" is as much an inaccurate comparison as "brains are like computers", the latest in a long line of similes which have in the past included telephone exchanges, hydraulic systems, and mechanical gears). Computers also lack

context, whether it be global and objective or local and subjective. Attempts to generate reliable high-quality metadata by machine without human intervention continue to fail.

So, can the audience just use metadata created by professionals, and ignore what is generated by the amateurs and stupids? Unfortunately, the answer must be negative. There are just too many information resources requiring description, and professionals cannot meet the audience's needs on their own. Those needs are wide-ranging, and expectations of how they should be met are changing, not least because of the apparent empowerment of the audience to create its own metadata. The end-user is not usually concerned with the needs of others outside of their own peer groups, and is generally not interested in the finer points of the ambiguities of language and esoteric aspects of vocabulary control. Search engines seem to find what the end-user wants.

Can we rely on the metadata created by amateurs? Certainly not. Doctorow [2] makes some pertinent observations:

- People lie
- People are lazy
- People are stupid

We can therefore expect to find deliberate falsehoods, incomplete information, and unwitting falsehoods in user-generated metadata. But we may also obtain valuable and truthful metadata from users, as the following examples from the National Library of Scotland demonstrate. These are taken from Flickr, where the National Library of Scotland has been investigating the potential for encouraging access to its digital collections by exposing sample images [3]. Flickr is a social networking service, allowing contributors to publish and share digital or digitised photographs, and add descriptive metadata and tags. Tags are keywords assigned to an information resource to help describe it and allow it to be found by browsing or searching; tags are therefore another form of metadata. As is typical in such services, users are encouraged to make comments on what they find and add their own tags, and comments are displayed alongside the image and its metadata. In some instances, the comments themselves constitute unstructured metadata.

flickr from YAHOO!

Home The Tour Sign Up Explore

Tommy enjoys possession of newly captured Hun trench

British soldiers at the old German Front Line, during World War I. In front of a mound and standing in a network of trenches are groups of soldiers, mostly smiling and laughing. They are all wearing large ponchos and the ground is very muddy. One soldier is pointing to a sign which says the 'old hun line'. [Original reads: 'OFFICIAL PHOTOGRAPH OF THE

Figure 1: Partial screen-shot of a photograph added to Flickr by the National Library of Scotland.

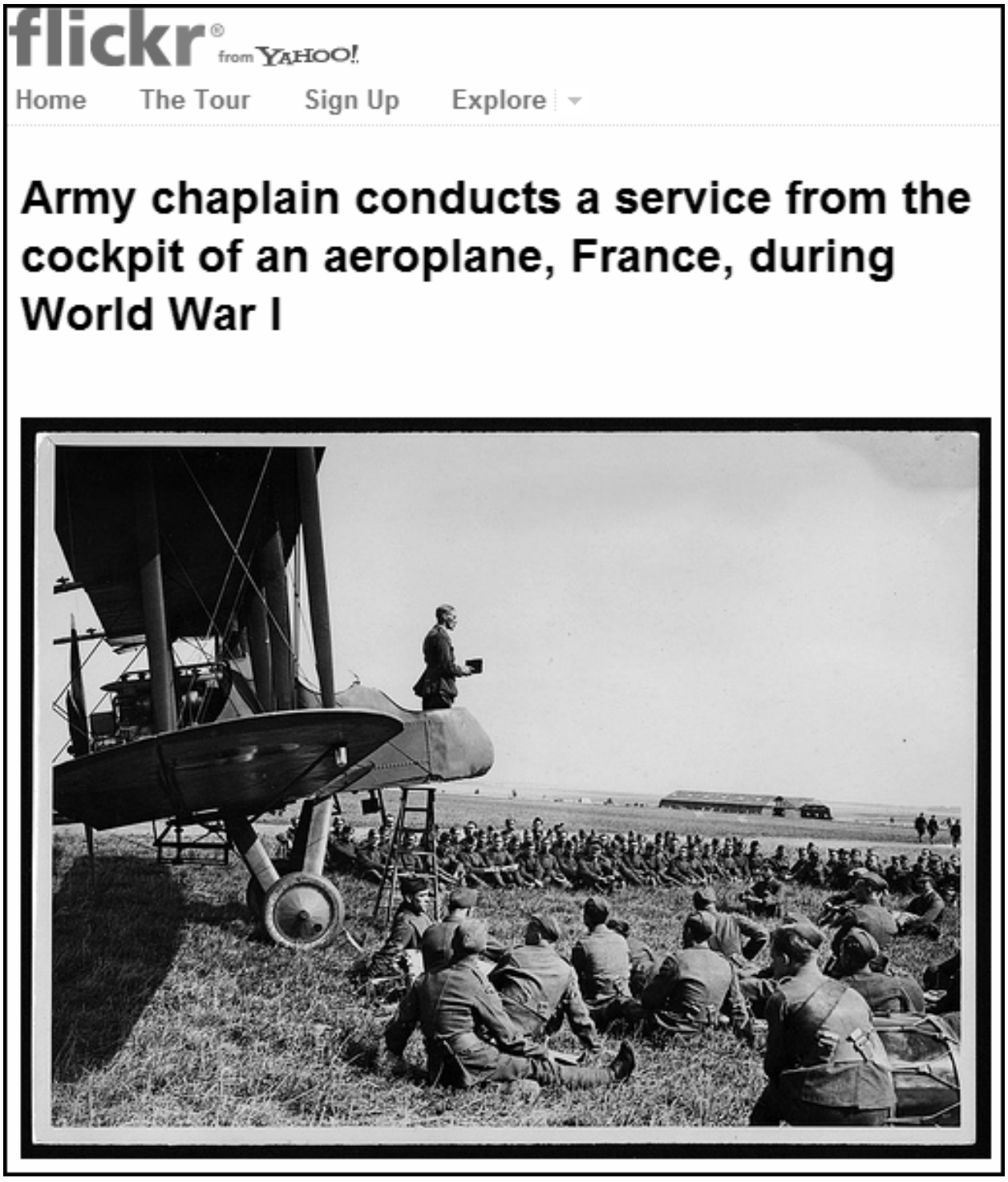

Figure 2: Partial screen-shot of a photograph added to Flickr by the National Library of Scotland.

Figure 3: Partial screen-shot of a photograph added to Flickr by the National Library of Scotland.

These photographs attracted the following comments from users (edited to correct punctuation and remove some superfluous words). For the photograph in Figure 1:

"I really can't believe that I am seeing the photo of my Grandad on the internet. I have had the photo all my life, and there were in fact two photographs taken at the same time only slightly different. My grandfather is the young man in the trench at the back. His helmet has a dent in it which was made by a German bullet. He was in the Royal Warwickshire Regiment, I believe it was the 16th Battalion. He told me that all the men in the picture were killed; he was the only one who survived. His brother was in the Royal Artillery, and they both went through the war, with my grandfather joining at fourteen."

If this information can be verified, it is a significant addition to the metadata for the photograph. The name of at least one person, and the name of the regiment, could be added as entry-points for retrieving the photograph. Knowledge of the existence of another photograph may help with collection development and curation.

Another user commented on the photograph in Figure 2: "For some reason this photo looks to me like it is of a chaplain holding a service." The National Library of Scotland replied: "Well spotted – it's the wrong photo with the metadata or vice versa. I'll look into it and sort it out. Thanks for your heads up!"; and subsequently "Sorted now, I hope." So, even information professionals ("national" information professionals, no less) can make mistakes. Collaboration with information amateurs is clearly beneficial.

A user comment about the photograph in Figure 3 said "That's an SE5 or SE5a, note the top wing mounted Lewis gun on a sliding Foster mount". The National Library replied "Thanks. We've tagged the image with this info now.", and the user responded "You're welcome." Again, combining professional and amateur metadata has improved the service. And presumably user perceptions of the quality and utility of the service are enhanced by such interactions.

In these examples there is no evidence of lying, laziness or stupidity – but that does not mean that all amateur metadata is reliable. Professionals must double-check that the user-supplied information is true if the integrity of the service is to be maintained.

How reliable is the metadata created by us professionals? Sometimes we get it wrong by mistake, as the example shows. And metadata is not likely to be comprehensive and complete outside of national cultural and memory institutions. But we do not often lie, and in general the metadata to be found in archive, library and museum records will be reliable. Although we are not lazy, we are only paid to be professional during working hours.

What about the reliability of machine-generated metadata? Computers definitely do not lie, unless programmed to do so, and they don't make mistakes, unless programmed to do so. They are not lazy, either, and are capable of working 24 hours a day, 7 days a week, 365 days a year. But they are very, very stupid, and there is potential for lots and lots of unwitting falsehoods, depending on the software.

If information professionals, amateurs, and stupids are to contribute effectively to the creation of metadata, we must find a way of exploiting the strengths of each group and minimising the

weaknesses. Or to put it very simply: people (amateurs and professionals) supply the brains, and computers do the work. This approach underpins the current development of the Semantic Web.

Semantic Web

The Semantic Web is based on the idea of making the simplest possible metadata statements and then processing them by machine to synthesise or infer new metadata statements and build up more complex metadata structures. The simple metadata statements are called triples and consist of three parts: the resource being described (called the subject), a property or attribute of that resource (called the predicate), and the content or value of the property or attribute (called the object).

E.g. “This article” - “has creator” - “Gordon Dunsire”

Triples are represented using a data model called Resource description framework (RDF). “To facilitate operation at Internet scale, RDF is an open-world framework that allows anyone to make statements about any resource ... RDF does not prevent anyone from making assertions that are nonsensical or inconsistent with other statements, or the world as people see it.” [4] This feature of RDF is often known as the AAA slogan: Anyone can say Anything about Any topic. The computer programmes that process triples to support metadata applications must therefore include methods to deal with conflicting sources of information.

One way of supporting verification is to create metadata about the triples (metadata about metadata), describing what agent made the statement, and when it was made. Knowing the source of the statement allows it to be categorised as professional, amateur, or "stupid"; knowing the age of the statement can help to determine its current accuracy. It is easy to create triples about triples in RDF because Anyone can say Anything about Any topic. The method employs a process called "reification", during which one RDF triple becomes the subject of another RDF triple. Using the example above:

(“This article”- “has creator”- “Gordon Dunsire”) – "has creator" – "Gordon Dunsire"

In other words: "I said that I created this article". If it can be verified that I am an information professional, the original triple is likely to be reliable. In practice, the meta-metadata might be:

(“This article”- “has creator”- “Gordon Dunsire”) – "has creator" – "National Library of ..."

Similarly, the currency of the original triple can be indicated by:

(“This article”- “has creator”- “Gordon Dunsire”) – "has creation date" – "2009"

In other words, the two new triples combined say "The National Library said 'Gordon Dunsire created this presentation' in 2009".

In practice, the subject and predicate of a triple are referenced by a machine-readable identifier rather than human-readable labels. The object of the triple can be a label, but an identifier can be used if the label is taken from a controlled vocabulary. In RDF, the identifier must be of a particular type: a Uniform Resource Identifier (URI). A URI can have a similar appearance to the familiar

Uniform Resource Locator (URL) used to identify and locate Web documents, although other forms of URI are permitted.

The URIs created for a specific set of subjects, predicates or objects usually start with the same URL-type string of characters, known as a base domain and similar to the top-level or root domain name of a website; e.g. "mywebsite.isp.uk/". The resulting set of URIs, together with the labels, definitions, scope notes, etc. of the subjects, predicates or objects being identified, is known as a namespace.

Namespaces can be used for another way of identifying the source of RDF triples. Organisations are familiar with the use of domain names to associate their "brand" with their website by including an acronym or short name in the URL. This helps users to remember the URL or identify it when using a search engine, and is so successful that fake but similar-looking URLs are still the basis of "phishing" scams and identity theft. In the same way, including an organisation acronym, short name, or brand in the base domain of a namespace ensures that every URI in the namespace carries an indication of the source of identification. This technique is already used in several namespaces maintained by library organisations.

For example, the Dewey Decimal Classification (DDC) uses "dewey.info" as the base domain, resulting in a URI like http://dewey.info/class/330/2009/08/about.fr [5]. Similarly, the Library of Congress uses "id.loc.gov" as a base domain for namespaces for its authority files and other controlled vocabularies. These include Library of Congress Subject Headings (LCSH): "http://id.loc.gov/authorities/sh85040850#concept" is an example URI [6]. Vocabularies from RDA: resource description and access have draft URIs such as "http://RDVocab.info/termList/RDAContentType/1013", which identifies a term from a list of bibliographic resource content types [7].

This provides an initial at-a-glance verification for information professionals and amateurs that the namespace is maintained by a professional agency, but that is a secondary benefit. A URI is primarily intended for consumption by information stupids who cannot understand "brand", so machines have to be told which are more trustworthy. This is essential if a computer programme is to avoid processing false information when creating machine-generated metadata, as happens when a collection of RDF triples is used to infer new triples through semantic analysis. For example, the triples "B" – "has parent" – "P" and "S" – "has sibling" – "B" are used to infer the triple "S" – "has parent" – "P". Of course, the computer does not understand what any of the triples mean; instead, the inference rules have to be specified as part of the programme. RDF is designed for a distributed metadata environment, where triples for processing are collected from multiple namespaces and "triple stores" of statements about specific instances of information resources. Large numbers of triples are required for useful applications, such as wide-area, cross-domain information retrieval, so it is inevitable that inference collisions or clashes will occur as a result of accidental and deliberate mistakes in the collected triples. Branded URIs offer possibilities for resolving such collisions, as well as indicating the quality of the metadata to the end-user. Thus the programme could instruct the computer to cross-check the URIs against a graded list of trustworthy base domains, calculate a quality rank for each set of linked triples leading to the collision, and favour the inference with the highest ranking. A quality rank might be determined from a trust grade for the base domain combined with the date of the triple. A trust grading might take into account factors such as the standards compliance of the organisation maintaining the namespace, the method used to create

instance triples, and the aims of the organisation; in fact, the same factors used to assess the quality and reliability of catalogue records, archival descriptions, and museum documentation.

Reification thus offers an explicit way of determining who made a metadata statement in the Semantic Web and when, while a branded namespace offers an implicit indication of the source of the metadata.

Conclusion

So the answer to "Who said that?" is likely to become very important if professionals, users and machines are to collectively make effective use of the Semantic Web in creating, developing and maintaining metadata for information retrieval. Archives, libraries and museums are held in high esteem as guardians of cultural heritage. We are trusted organisations, and our professional methods produce trusted output. We should try to get our trustworthy metadata into the Semantic Web as quickly as possible and make sure that we are identified as the source, explicitly by using reification and implicitly by using branded namespaces. Our metadata can be used to resolve disputes with that from less-trusted sources such as computers and end-users, as well as helping avoid unnecessary duplication and encouraging the creation of high quality metadata by non-professionals by exposing the output of professional practice. We should be proud to identify our metadata: We said that!

References

[1] WordNet entry for metadata. Available at: http://wordnetweb.princeton.edu/perl/webwn?s=metadata

[2] Doctorow, C. 2001. Metacrap: Putting the torch to seven straw-men of the meta-utopia. Available at: http://www.well.com/~doctorow/metacrap.htm

[3] National Library of Scotland's photostream in Flickr. Available at: http://www.flickr.com/photos/nlscotland/

[4] W3C. 2004. Resource description framework (RDF): concepts and abstract syntax. Available at: http://www.w3.org/TR/rdf-concepts/

[5] Dewey.info. Available at: http://dewey.info/

[6] Library of Congress. Authorities & vocabularies. Available at: http://id.loc.gov/authorities/

[7] NSDL Registry. RDA content type. Available at: http://metadataregistry.org/vocabulary/show/id/45.html