David Giaretta and Mark Dunkley

Authenticity Capture Prototype

slide 1 This work concerning the Authenticity Capture Prototype has been done by David Giaretta in collaboration with Mark Dunkley.

slide 2 Other presentations explain the reasons why authenticity is very important in the context of digital preservation. Here, a tool is presented which may allow the capture of different types of evidence from a number of sources. To reach this objective, it is made in such a way as to be customizable and flexible, without being prescriptive, as many ways exist to capture the various different types of authenticity.
This tool is essentially a framework into which the Authenticity Protocols (APs) and the Authenticity Steps (ASs) can be plugged, in addition to the various types of evidence shown in the other presentations, may be inserted.

slide 3 The tool implements the Authenticity Model.

[This slide has not been presented at this point but at the end of the speech.]

slide 4 The tool is a framework that can be built on in order to add capabilities through plug-ins. The developers started by capturing mostly textual information but it is also possible to add attachments of various kinds.

slide 5 They also wanted to allow a number of different user roles and user cases. The Project Administrator, in the process of capturing information evidence for authenticity, creates a project which will contain the information that is captured or, indeed, he or she may import some existing provenance and evidence which will be appended to the resources in the process of the chain of custody.
Another important role is the person who captures the evidence: he/she must be authorized in some way to add evidence to the existing collection, so he/she has to register, have a user profile, create various types of evidence and insert them into the collection of evidence that is being built up, so as to sign out.
The third role is that of a researcher or a general user who wants to examine this collection of evidence and make some sort of judgment about whether he or she really believes that it has a sufficient ‘degree’ of authenticity to be useable. A variety of ways of browsing that evidence, and of using and exporting it so as to be used in other tools has also been allowed.

slide 6 One item which is called Digest, has been inserted into the framework with the aim of answering the question of the great quantity of flexibility necessary, in terms of the information that is captured and of the various APs and ASs.

There is the need to be able to say that the captured evidence has not been tampered with; therefore each step has to be signed by somebody to certify that it has been checked or examined, so that person has to be able to sign that it was not just a fake entry.
Cryptographic hashes are used at various times in the process so that an Information Capturer can certify to the fact that he or she has added information to the collection of evidence and then, at the end of process (because that person might have included many AS) when he signs out of his/her session, another hash is created, which makes sure that we can check that the collection of evidence which has now been extended, has not been tampered with.
A whole collection of hashes allows one to check that the collection of evidence has not be tampered with at any time and makes sure that the person who has added the evidence is really who he/she is believed to be.
A variety of hash techniques can be used. The important point is that the Information Capturer essentially takes the whole of the message, the whole of the collection of evidence, and it is then impracticable to substitute that hash, to modify the message without showing up when the hash is recomputed, and it is impracticable to say that he or she has not inserted that evidence.
Of course one question arises as to where this hash should be kept; there are number of projects which have been addressing this issue and the simplest way would be to take the hash that has been created from the latest collection of evidence and put it somewhere in a public place so that it cannot be tampered with easily. However, the hash is also contained within the collection of evidence.

slide 7

[An example in the presentation shows that a user (an Information Capturer) has a name, and that this name, the hash algorithm**,** the hash value, and other information are contained within the evidence, encoded, in this case, in XML.]
At any point in the collection of the evidence, at any AS, the Capturer may want to express a degree of confidence that he has checked it sufficiently. Due to the fact that by simply looking at data, it is almost impossible to check every single value, the user is therefore allowed to give a percentage of confidence; clearly a high judgment on the part of the person who has captured the evidence.

slide 8

An XML schema is being developed to encode this information which is based on the Authenticity Model.
CASPAR researchers are in the process of defining an XML schema to capture the information. The evidence that is being collected is based on the Authenticity Model which is presented in the other statements.
As Maria Guercio underlined, the flexibility of this framework is extremely important also in the definition of the terminology to use. In fact, there are many other sources and many other ways of capturing authenticity, such as the PREMIS model and its encodings, in addition to many different kinds of metadata; so being able to insert these other sorts of evidence into the container that has been created, is really important too.

slide 9 One example concerning the user experience may better explain what has been said above.

A user may enter into the tool, identify him/herself, select a project, i.e. some pieces of data to which evidence is going to be added, select the sort of things that are going to be done and then start to go through the different sorts of APs and ASs and parts of Preservation Description Information (PDI) that are going to be inserted. For instance, the user might select the ingest protocol, the sort of data set steps within that protocol and a particular type of PDI to add; at this point, as is discussed in the Authenticity Model, there will be some recommendations, which have been created by the author of the AS and the AP, and the user can then put in just simple text, or something that should also be cut and pasted from other sources of information, or something that can be inserted as an attachment. (The percentage of the Information Capturer's confidence that this is the source of the information to which authenticity evidence is being attached is shown on the same page).

A variety of visualization plug-ins are available on the tool; for instance, at a later date, another user may come along, look at all of this evidence and find some visual clues that express how confident people were of the sort of evidence that was captured and inserted .

slide 10 A case study shows that the developers wanted to allow users to design the APs and ASs; as part of that, a statement in policy and a recommendation should be inserted. In designing that protocol, a number of events can be added. Finally, also the details of the protocols and the individual steps should be inserted.

slide 11 This protocol can then be ingested into the tool, and a variety of formats can be supported.

slide 12 Another example shows something that may happen quite often: a transformation of the data.

A recommendation policy might be included in the protocol which states, for instance, that "for the received data file to be deemed as sufficient quality to support data analysis it must have been successfully transformed into standard IIWG (i.e. a particular scientific data format), the use of the processing software must be recorded".

The steps may be, for example: details of the transformation (the version of the software, the source, time and date, who is responsible, the reasons for believing the software is reliable) and the evidence about the transformation. All of this type of information can be captured in the tool.

slide 13 Another example concerns the types of checks that have been performed in the process of transferring the data from a local machine into a persistent storage (e.g. the final validation of the IIWG file as it is transferred from the users local

storage to a more persistent storage); the provided recommendation concerning the criteria, and the types of information and evidence that will be captured. (The presentation is followed by a video that shows the tool in action.)

slide 3 The tool that has been built, is a prototype; at the moment it is available with the collection of other CASPAR software but it is also a standalone tool which can be used in a variety of situations, as it is a general tool which is relevant for all certification, and it should be extremely useful to everybody who is concerned with the authenticity of any sort of data. Plug-ins will be available which can import existing sorts of information and extract information from existing data files, and the evidence that is collected will be 'queryable' in a variety of ways so the users of a particular digital object will be able to get a good feeling about how to judge the authenticity of that particular digital object. This sort of information should be very helpful for an NI repository, for example, which needs to be audited and certified.