Project InterParty: from library authority files to e-commerce

Andrew MacEwan
The British Library

Origins and overview of the project

When I was asked to give a presentation on the work of the InterParty project the working title that first occurred to me was "From library authority files to e-commerce". This seemed to capture the thought that we were pushing boundaries by applying the familiar principles and benefits of library authority files to the management of e-content by working in co-operation with publishing and trade sectors for whom authority control is not a familiar notion. On reflection I think it is a more accurate reflection of the key drivers of the project to reverse the caption to read "From e-commerce to library authority files". For in fact this is a project initiated by people working in the trade sector who have identified a "problem" and seen part of the solution as pre-existing in library authority files. I hope to show in this presentation that, although InterParty is not a library-led project, it does offer to open up potential new partnerships that would greatly benefit authority control work in libraries.

What is the InterParty project? It is a project, funded by the European Commission, which aims to develop a mechanism that will enable the interoperation of identifiers for "parties". The term "party" is simply a useful term with which to draw together the disparate types of identities responsible for the creation of intellectual property or "content", such as authors, composers, performers (including groups), producers, directors, publishers, collecting societies and even libraries. The project brings together partners from the book industry, people involved in rights management, libraries and identifier & technology communities; all of whom share a common interest in the accurate identification of "parties" in relation to "content" for varying purposes.Specifically the partners are: EDItEUR (the co-ordinating partner, a European organisation founded by European Federations of Library, Booksellers and Publishers Associations to co-ordinate the development, promotion and implementation of EDI in the books and serials sector); the British Library; the Royal Swedish Library; IFLA; Book Data (a leading supplier of bibliographic data to all sectors of the book supply chain in the UK); Kopiosto (a leading copyright agency operating across all the creative media sectors in Finland). US-based partners, not funded by the EC, are: the Library of Congress; OCLC; the International DOI foundation and CNRI (Corporation for National Research Initiatives).

The original idea for the project was an outcome of another EC project <indecs> (Interoperability of Data for eCommerce Systems). <indecs> was similarly concerned with transactions in content and how they can be effectively controlled in a web environment. In its analysis <indecs> proposed that descriptions of content, transactions and descriptions of rights are all inextricably linked, and recognised that accurate descriptions of content are the core on which the rest is based. Consequently the key outcomes of that project were the definition of a generic data model and the promotion of mapping to that model from specific sector models. The <indecs> model

is the foundation for the ONIX Data dictionary that is now the international standard for representing and communicating book industry product information in electronic form.

Central to the <indecs> view on the need for interoperability of metadata about resources was the recognition that a critical part of that metadata is accurate identification of authors, creators, etc. The project left this as unfinished business in the form of a proposal that further work should be undertaken to develop a system for linking existing person identifiers through a Directory of Parties. It is this proposal which InterParty is taking forward and which will complete the picture of interoperable metadata required to support discovery of resources, discovery of rights ownership, negotiation of agreements, payment of royalties and other potential applications. These are substantial goals!

So it is important at this point to emphasise that the InterParty project is aiming at this stage to deliver no more than a working Demonstrator - an Alpha system - not a working network. The project is funded to run for only 12 months and it will effectively aim to provide a proof of concept that can be demonstrated to potential members of a future live network. The key goals for the project are the specification and building of the demonstrator system (this will simulate the interoperation of a network of participating databases); and the development of a business model and governance proposals for a real-world implementation. Supporting these two key goals will be an analysis of existing data models, such as those already underpinning the databases of author licensing agencies and library authority files; the development of an interoperable party metadata model; and an analysis and resolution of any potential privacy and security issues which might arise.

That is the overview of the project in terms of its origins, aims and goals. It is time now to look in more detail at the path that leads from problems in e-commerce to library authority files.

The InterParty analysis

The starting point of the InterParty proposition is the recognition that there are already plenty of existing databases containing metadata about people and organisations and these serve to accurately identify "parties" within their own context. At the present time most of these databases are entirely independent of one another, following different approaches to identification and involving different schemas and formats. Even within sectors differences in standards are an obstacle to sharing metadata - a problem very familiar to the library sector. We can also appreciate from a library perspective the historical path which has led to independent "data silos". In the past the need for such a level of inter-connectivity was neither apparent nor easy to achieve.

Today the growth of the web has highlighted the need for metadata which can "travel" across these standards and systems barriers. Again the InterParty analysis concerns barriers to communication across many different sectors with much greater diversity between them than the barriers that exist between libraries internationally, but the arguments are the same. Sharing metadata about parties would improve efficiency, it would improve effectiveness of communication and it would support navigation between domains and services on the web. The difference is that the level of barriers that exist between sectors in terms of lack of standardisation is much greater and this in turn links to the different business requirements of different sectors.

From a library perspective the key business requirements are simply an extension of our standard requirements for authority control: access to new sources of metadata with which we can enrich and improve the quality of our own authority files. Ready access to useful additional sources of data already held elsewhere in the wider trade sector would also provide a quicker, potentially more efficient, means of resolving identification problems: is author X the same as author Y? It is easy enough to see that benefits to libraries would be similar in other sectors, for instance in terms of service to end users. A retailer would be able to better support user requests for information on all the recordings of a particular composer, e.g the recordings of John Williams - but which John Williams? Such requirements are not commercially critical and so from the InterParty perspective they are noted as requiring a reasonable degree of certainty of identification.

By contrast any organisation involved in rights management will require access to sources of metadata which will support business transactions. In the most extreme case this may involve trusting the metadata in an authoritative record for a given "party" with a view to using the data to ensure a payment is made to the right person at the right address. Such requirements involve a very high degree of certainty of identification.

The InterParty proposition is that these disparate business requirements nevertheless converge on a common need for accurate metadata to support the identification of parties. Benefits in terms of quality and efficiency would be gained for members of a cross-sector network because there is a common functional goal in all the databases: the unique identification and disambiguation of parties. Although the goal of unique identification takes on a sharper degree of importance for commercial purposes it remains a common qualitative goal for all sectors.

Having established a common benefit which can be derived from interoperation how does InterParty intend to resolve the inherent problems? The fundamental aim of InterParty is to develop mechanisms that will link the existing, disparate databases currently used in different sectors to record and control the identification of parties. InterParty will therefore be a "membership" network of InterParty members or IPMs and these will comprise organisations with metadata to share, and identification schemes to support. Specific membership criteria will need to be defined as part of the Governance model. Members will join InterParty because they perceive a common benefit from interoperation at the very least in terms of access to "common metadata" held by other members to improve the quality of their own data. Potentially, the development of links between different databases will also support automated machine-to-machine "transaction".

Members will be able to derive new identities on their own databases from other IPMs but InterParty itself will not originate new "party" records. Individuals and organisations ("parties") will only be identified within the InterParty network if information about them appears in one or more sets of data created or held by an InterParty member.

The InterParty network or system will provide a "resolution service", a single point of access to the multiple databases on the network. Each database will comprise its own "namespace" – the metadata context within which entities are uniquely identified. Each "namespace" on the network will make available to the network a specified subset of "common metadata" sufficient to disambiguate each identity within its own namespace, excluding where necessary any data that must be restricted for reasons of confidentiality.

To define the "common metadata" required InterParty draws upon the definition of metadata used in the <indecs> project: "an item of metadata is a relationship that someone claims to exist between two referents". For instance a relationship between a name and a variant form of that name, or between a name and a date of birth. Note that a key part of the definition focuses on who makes the claim of a relationship. In library authority files source of information can give crucial validation to a record, e.g. a letter from the author. Some databases within the InterParty network may be able to provide more of this kind of validation or "authority" than others. All the member databases will already express many such relationships. The InterParty network will add a new layer to these by enabling new relationships to be expressed and recorded as InterParty "Links", e.g "Person X in Namespace A is the same as Person Y in Namespace B".

Affirming this new level of metadata will require effort and judgement. Although it may be possible to automatically generate or propose many links on the basis of algorithms this is not being developed for the demonstrator system. Potential links will also be identified in the course of each IPM using the network to derive information to confirm or validate the relationships in their own databases. By recording the discovery of relationships between identifiers in different namespaces InterParty will ensure that the effort is not wasted but is made available to all on the network for future reference.

To make this new information available will require a basic format in which an "InterParty Link" can be expressed. In principle the link information could be held locally within IPMs as part of the common metadata set or it could be held centrally in a separate InterParty Link Database. For the purposes of the demonstrator the project will simulate the latter model. This will require further analysis before confirming the approach best suited to a scaled up working system.

Common metadata and Public Identities

The main task being addressed for the demonstrator is the definition of the metadata required for the InterParty model. It is on this issue that InterParty begins to cover ground familiar to library authority files. The fundamental requirement is that members will need to provide access to sufficient metadata to achieve disambiguation between parties with shared or similar attributes, and also collocation of the same party when they have different attributes, e.g. John Williams the composer is distinct from John Williams the classical guitarist who is the same as the John Williams who formed the group Sky. How much metadata is sufficient will depend on the context. If a given database only contains a record for one John Williams then the name itself is a unique identifier and no other defining metadata is required (although it may be useful to record some in the event of future additions to the database).

Because InterParty is potentially dealing with databases that contain metadata about people that may be commercially sensitive or just private it has defined the "common metadata set" in terms of information which is in the public domain. By focussing on this subset of information it becomes clear that actually what we are dealing with in identifying parties is, in the case of real individuals, a construct of the real persons underlying them. This construct InterParty has termed the Public Identity.

An individual person may have one or more public identities, most obviously in the case of authors using one or more pseudonyms. The notion of a "public identity" is similar to the concept of a "bibliographic identity" which has been defined as a key

entity in the current draft of the FRANAR data model for name authorities. The question of whether someone has more than one "Public ID" is a matter of "functional granularity". Although pseudonyms provide a useful example of the concept a Public ID is not the same as a name since more than one name may be associated with the same public identity. Sometimes, relationships between Public IDs are not public, but become so, for example, Ruth Rendell and Barbara Vine. And sometimes, two or more people may share the same Public ID, for example, Nicci Gerrard and Sean French writing as "Nicci French". This, too, may or may not be publicly known.

Key definitions

To clarify this further here are some of the key InterParty definitions around the concept of public identity.

Party: An individual or organisation involved in the creation or dissemination of intellectual property
Public Identity: An identity that is associated with and is used publicly by a party (or a group of parties)
Public Identity Identifier (PIDI): An identifier assigned to a public identity by an IPM and designed to be unique within the domain of that IPM: a PIDI may be a number, or it may be a controlled form of name (eg in a library name authority system)
InterParty Link: An assertion about a relationship between two PIDIs in two different IPM domains - ie, between two public identities

InterParty is concerned with asserting relationships between Public Identities in different namespaces. Within the InterParty network each Public Identity will require a Public ID Identifier (PIDI) which will comprise a combination of identifier for the namespace and a unique identifier within that namespace. InterParty Links will express relationships between PIDIs in different namespace domains in the InterParty network. Each PIDI will represent a set of "common metadata" which the IPM owner of that namespace is prepared to make publicly available over the network.
What should be provided as "common metadata" will depend on the agreement of the InterParty members and will depend, critically, upon their willingness to share data currently available only to their own users with a wider network of "foreign" users not related to their core business or purposes. The minimal requirement is for a practicable set of data elements that is sufficient for the purposes of disambiguation and which can be regarded as in the public domain. At this stage the project is proposing a set of data attributes for the common metadata set and validating them with potential InterParty members through a combination of questionnaire and workshop sessions.

Proposed Common Metadata Set

The current list of data elements comprises the following:

PIDI

The unique ID, comprises Namespace:identifier
Identifies the IPM and the Public Identity

Must be persistent, though the associated metadata will typically change

Name

The name(s) by which a Public Identity is known
Name types may include: Preferred (standard) form; Known variants; Former names – with dates

Events

Significant events and their dates, and places where applicable
eg Birth, Death, Incorporation (for a corporate Public Identity)

Works

Works which with Public Identity is associated, represented by title accompanied by date & role of Public Identity if known

Roles

Roles typically performed by the Public Identity or spheres of activity – not just directly in relation to works
eg novelist, conductor, footballer, politician – with dates where appropriate

Relationships

Relationships with other Public Identities
e.g. has collaborated with X, has illustrated books written by Y, in same band as Z

Affiliations

Formal or official positions held by Public Identity
e.g. Professorships, and memberships of organisations, societies, etc

InterParty Links

Access to the Links is key element of Common Metadata

There are some conceptual difficulties with relating some of these attributes to a Public Identity as defined. Many of the attributes relate more properly to real persons than to public identities as defined. Such attributes may be considered to relate to the public identity insofar as they have been made publicly available in the course of that "identity" releasing a work of intellectual property. This can be problematic with regard to attributing dates of birth, etc. to pseudonyms considered as discrete public identities. Normally the extension of an attribute from the underlying real person to one or more of their public identities will be a simple transference but in extreme cases a pseudonymous identity may take on a life of its own. Nicci French, whom it has already been noted is the Public Identity representing the collaborative output of two real persons, has gathered some real-world attributes in a recent advertisement for their latest thriller: "Nicci French's bestselling novels are The Memory Game, Killing me Softly, etc. *She lives in Suffolk*"!

Once the Common Metadata set has been agreed we will need to define rules and appropriate format conventions (currently being defined in a draft XML schema). The more standardised the Common Metadata (in terms, for example, of controlled "values") the higher its value – but the higher its cost. The extent to which the "common metadata" will need to adhere to common forms of semantic or syntactical expression will depend on some, as yet undecided, issues concerning a real-world implementation of the system. If the primary use of the network is direct human access and interpretation of data on a case-by-case need then only limited standardization will be required. If large-scale algorithm-based linking operations are to be run there may be a requirement for more standardized data.

Finally it cannot be expected that all IPMs will be able to provide metadata for all the proposed categories. Currently, the only mandatory elements are expected to be the

PIDI and at least one name. This is the minimal practicable data on which links will need to be based but clearly more data than that will be required to inform either human or algorithmic decisions about links.

InterParty Links

Let us now look in more detail at the InterParty Links. As already indicated this is the added-value category of metadata that the InterParty network proposes to offer.
An InterParty Link is the assertion of a relationship between two Public Identities, represented by PIDIs. Any InterParty member (IPM) may propose a Link provided that they own one of the PIDIs that is being established in the Link. The link may then be endorsed or disputed only by the IPM that owns the other PIDI proposed in the Link. Any other IPM may add comments to the record but only the two IPMs that own the namespaces concerned may make or modify the assertion of a relationship.
The assertion of a link between two PIDIs is held in a single record. For the purposes of the demonstrator project, the relationships expressed in an Assertion will be restricted to "is", "is complex" and "is not". The record structure is defined so that other relationship values can be added in the future if required. For now other relationships, such as this company is the owner of that company, will be supported only within the databases of individual IPMs.
The relationships are being kept to the level of simple functional equivalence. So PIDI 1 "is" PIDI 2 asserts that PIDI 1 and PIDI 2 have a functional and reciprocal equivalence for the purposes of InterParty.
PIDI 1 "is not" PIDI 2 asserts that PIDI 1 does not have a functional equivalence with PIDI 2 despite appearances.
In order to keep the relationships simple a third type of complex equivalence has been defined to cover a variety of more complex situations that cannot fit into these first two categories.
PIDI 1 "has a complex relationship with" PIDI 2 asserts that PIDI 1 has a partial equivalence or complex relationship with PIDI 2 that is not necessarily reciprocal
This "is complex" relationship is designed to handle the different ways in which IPMs may hold records for public identities, parties and names in certain circumstances. For instance IPM A assigns a single PIDI for Ruth Rendell, with a note that Barbara Vine is a pseudonym of Ruth Rendell; but IPM B assigns separate PIDIs (i.e. separate records) for both Ruth Rendell and Barbara Vine (with or without an internal assertion between them). It cannot be said that IPM A's Ruth Rendell/Barbara Vine "is" IPM B's Ruth Rendell, although there is a relationship. This is expressed as "complex".
There are numerous other circumstances where it cannot be assumed that all IPMs will take the same approach to identification – or even be aware there is an issue. Cases of an author using multiple pseudonyms or two parties combining under the guise of a single pseudonym, as in the Nicci French example will all tend to fall into this category when different IPMs capture and describe these public identities in different ways. It is not proposed to define all the relationships covered by "Complex" any more precisely at this stage of the project but examples of complex relationships will be included in the demonstrator system.
Since the assertion of these three types of relationship will involve actions by different IPMs over time the Link records will also need to record the current status of

the assertion being made. The status of a link will relate to how it is established and to what degree the two IPM owners have been involved. There are 4 status types

"Proposed"
The relationship has been asserted by one IPM owner only
"Authorised"
Concurring assertions have been made by both IPM owners
"Disputed"
Assertions have been made by both IPM owners but they do not concur
"Inferred"
Generated automatically based on inference from 'is' relationships only

Although the primary mode of making and editing links on the demonstrator will be manual it was felt useful to build in a further category of automatically generated links. These are 'inferred' links that can be derived from assertions of the type, PIDI 1 'is' PIDI 2. Where a PIDI has become involved in more than one link of this kind it will be possible to infer further relationships. So where PIDI 1 'is' PIDI 2, and PIDI 2 'is' PIDI 3, the system can infer that PIDI 1 'is' PIDI 3.
The current draft outline of the Link record which is proposed for the InterParty demonstrator contains the following elements:

Draft Outline of a Link Record:

Link ID
Unique identifier for the Link Record
PIDI 1 (Namespace:Identifier)
Identifier of Public ID
PIDI 2 (Namespace:Identifier)
Identifier of Public ID
Link relationship
Code indicating nature of the relationship asserted, i.e. 'is', 'is not' and 'is complex'
Link status
Value indicating level of trustworthiness of the link, ie. 'Proposed', 'Authorised', 'Disputed', 'Inferred'
Link method
Manual or automatic
Link creation/update timestamp
Timestamp indicating when the record was created or last updated
Owner Assertion composite
A group of elements which record each Owner IPM's assertion about the Link, including
-Owner ID
-PIDI owned
-Owner assertion – used to set up/amend Link Relationship type above
-Assertion comment – notes field
-Asserted by – name of individual
-Assertion timestamp

Comment composite

A group of elements to allow other IPMs to add further notes/comments to the record without directly affecting status of the assertion

Although there are quite a number of data elements listed the intention in the functional specification is to make creation of a link as simple and effortless as possible, with automated defaults and simple routines for selecting and entering the PIDIs into the Link Record. Further considerations for processing links include a facility to automatically alert the IPM owner of the second PIDI to the presence of a new link whenever an IPM initiates a link. Creating a link will always trigger the start of a validation process. Only when both owners of the link have asserted the presence of their PIDI in the link will the status of the Link become fully authorised. To allow flexibility it will not be mandatory to complete the validation process but not to do so will weaken the authority or trustworthiness of the link. It will also be possible to reverse the authorisation if required, for instance if new information calls it into question.

Finally the links themselves will be retrievable via their record control numbers (Link IDs) or via the PIDIs within the links. It is assumed the value of such searches will grow as the universe of proposed and authorised links grows on the InterParty network. Further uses of such control numbers within the domains of individual IPMs are a matter for speculation and are not a part of the InterParty system. But it is possible to imagine InterParty Link IDs gaining a value in their own right as reference points to a network of metadata concerning an individual Public Identity.

Conclusion

At the beginning of this paper I emphasised that the InterParty project is a demonstrator project due to complete on a short timescale - by mid-2003 - in order to offer a proof of concept simulation which will illustrate the potential value of a real world implementation. The key questions will be answered after the demonstrator is complete. Who will want to join InterParty? If a network is established will members really want to invest time in creating and editing links, or will it be seen as just a search service?

It is likely that a real world implementation will have to address the question of automated, large scale production of links by means of algorithms in order to provide the InterParty added value as an early benefit. What remains certain is the level of interest in the problem that InterParty has set out to address. The basic benefits of authority control are clearly perceived as benefits that are needed outside the library sector. The goal for InterParty is to offer a realisable solution to the problem that is relatively cheap because it is based on interoperation and co-operation, not on the creation of a new standard. I see two main potential benefits to support the engagement of library authority files in an InterParty network. There will certainly be a benefit in terms of access to new realms of metadata that can enrich our own authority work. A further benefit may also derive from content producers and publishers using data (names or IDs) that is already linked to library authority files through the network. In the end, as is the case with all co-operative proposals, success is likely to depend on a few key players coming on board at the outset to make the initial investment.

References:
http:\\www.indecs.org