Making Information and Referral Resource Data More Accessible

May 31, 2014

by Eric Jahn, Data Architect
Alexandria Consulting LLC
commissioned by the Alliance of Information
and Referral Systems (AIRS)

Table of Contents

Preface

This document was commissioned by the Alliance of Information and Referral Services. Its purpose is to examine the existing information and referral (I&R) system and to propose ways to make it more interoperable, and fluid while making the resource data more accessible. Numerous sources were consulted, including multiple information and referral systems (in particular Chris Mattmann and Mark Perryman from 211 LA County), AIRS staff, information and referral software providers, Code for America's Open Referral project, and human services and general technologists. Their insights and advice were all very appreciated and helpful. A set of summary slides associated with this document are available at: http://tinyurl.com/q6jz8pc

Executive Summary

The existing AIRS information and referral network generally requires help seekers to call or search a resource website to obtain human services directory information. Increasingly, third party case management software and various applications (some mobile) are serving as intermediary tools in this process of matching resources to help seekers. Theses tools search across regions and provide an added layer of convenience or coordination to the process. An established and automated way to find and retrieve resource data does not exist. Using simple, existing technologies, information and referral systems can build components which allow third party applications to automatically access resource data, while preserving existing systems and data restriction.

The components include: a machine searchable directory of information and referral services, a simple gateway protocol for data acquisition negotiation, and a common methodology for publishing navigable linked resource data. These tools, when used in conjunction, automates access to the existing resource data. The agencies that owns the data (such as 2-1-1s and Aging and Disability Resource Centers) can control the terms under which the data is accessed and used to a fine grained level. The directory, gateway protocol, and linked data components would fundamentally improve the availability of resource data, effectively building a predictably accessible "system of systems".

Problems Identified In Current System

The following is a list of general observations of the existing information and referral system.

- The system works for people accessing help directly, but not for third party tools helping people. Automated tools can not yet scan large amounts of data across many information and referral systems to obtain the best information available, and to save the help seeker time.

- Interoperability is low.
  - Information and referral system providers and their implementations do not have a standardized and established way to send data amongst each other.
  - The existing AIRS XML Schema standard resource payload format is suitable for expressing chunks of resource data, but there is no standard methodology for how to send and receive these chunks of data between any two systems.

- The national level directory at 211.org is the only general directory of information and referral systems that exists. The directory however, is not suitable for access by software; it is only directly accessible by help seekers browsing the web. Allowing third party software tools to locate a specific I&R system serving a certain geographic area, will expedite the search process. Using information gained from the directory, these tools will then negotiate the terms of access and navigate the available data.

- The information and referral system should allow multiple code sets to coexist, such as 2-1-1 LA County's Taxonomy of Human Services, healthcare codes, government codes, subset, specialized, folksonomy, and even UN humanitarian aid code sets. This inclusiveness will enhance AIRS' status as a world class organization, meanwhile driving awareness of its specific, certified code sets. AIRS certification status could be highlighted in a directory. The directory should declare the taxonomies each system uses to encode its services data, so that consumers of the data can choose compatible and appropriate resource data sets, including specialized, folksonomy mappings, or subset taxonomies. A typical pattern for resource description could be comprised of a declaration of the taxonomy set, then specification of a taxonomy service code pertaining to that set. This isn't to say that a given taxonomy won't comprise most of the available services data, but expresses the need that the overall system can flexibly accommodate different vocabularies.

Underlying Interoperability Requirements for I&R

The following section enumerates requirements for interoperability within an information and referral system.

Interoperability Requires Software Automation

It is not practical or feasible for a human to sift through the sheer amount of data that needs to be analyzed from multiple data sources. For example, a software search tool might need to search across many different I&R provider agencies whose boundaries subdivide the desired search area. Or, the software search tool may need to examine only a subset of data within a specialized resource data set buried within a large I&R system. Automated tools can navigate these complex sources and cobble results together across organizational boundaries, ultimately on behalf of the help seeker.

No Piece of Data Should Exist in Isolation; Data Paths Should Lead to And From Each Datum

Regardless of whether a resource seeking tool starts searching at the national directory level, or deep within a resource data set, there should always be paths to navigate to other related data. Data linkages between related resources provide the ability to browse connected concepts, for example, nearby facilities, similar service offerings, etc.. This is Linked Data. It's just URLs pointing to other places on the web, that are expressed in a specifically structured way ("RDF", not HTML[1]). Linked Data is built from familiar things, like the Web, URLs, and common data formats, like JSON or XML, to represent it. Each RDF makes statements about resources, and in the case of the AIRS partners, making statements about

[1] but there are even ways to embed linked data into web pages. see RFAa or microcode

I&R resources (i.e. a specific service in the community). An example statement could be "an emergency shelter program [subject] is provided at [predicate] agency site 341 Main Street [object]".

They are called "Triples", because the have three parts: subject, predicate, and object. The object is actually within the predicate.

More examples:
Harvest House [subject] has a [predicate] food program [object]
Harvest House [subject] located in [predicate]ZIP Code 44567 [object]
Harvest House [subject] located at [predicate] Address 3371 Birch Street[object]

For example, if a resource listing mentions linkages to related or nearby resources (even if they point to another system outside the current data set boundaries), the resource seeking tool can then follow those linkages to their destination. These interwoven relationships between data will take time to establish. The centralized directory of information and referral services will also help third-party software to determine where to start looking for resource data at a higher level, if needed.

Independence and Resilience of a Federated I&R System

The information and referral system should be resilient to specific points of failure, or independent of outside organizations. For this reason, allowing search software to traverse linkages between the resource data (peer-to-peer) is preferable to having search software constantly start all searches at the national directory of I&R systems (top-down approach). By that same token, reliance on third party search aggregators to crawl, then publicize the data, while useful, compromises the self-sufficiency of the overall information and referral system. A federated approach to I&R information sharing will decentralize the network of I&R resource information, and can easily be overlaid with private search for additional benefit.

Data Instantly Becomes Stale Once Separated from its Origin of Aggregation

Resource data that is copied or exported from the original aggregating information and referral agency may become stale or outdated the moment it is moved by a third-party application to another data store. Future updates to the original aggregating I&R agency database would need to be polled again by the third-party application to keep current. Instead of storing immediately stale data, it's best for third-party applications to fetch the information again each time it is needed, directly from the source, whenever possible. This allows I&R agencies to maintain the integrity of the data, so that incorrect information is not disseminated to help seekers. This implies a sharing model of continuous real-time access to resource data, and not one of bulk data transfers and reliance on duplicate data warehouses. Tools providing cross agency searching and filtering of I&R resource data will need to be afforded query mechanisms to efficiently access only the pertinent data from its origin.

What Needs to Be Done by AIRS to Make Resource Data More Accessible?

Build a Machine Readable Directory of I&R Provider Agencies

- The directory would serve as an entry point to third-party applications with no other point of reference to begin seeking data.

- It could expand the comprehensiveness of 211.org to include all types of I&R provider agencies.

- The Directory could contain the following information on each I&R providing granted, ideally represented as the Linked Data (as RDF):
  - a URL to the gateway from which access may be negotiated and granted;
  - a listing of licensing, fees, or other contingencies attached to accessing the data at the Gateway URL;
  - the geographic/geospatial boundary of I&R, or a least a list of ZIP codes, or the equivalent; and a
  - a listing of the code sets used by each resource data set available.
  - (optional, but useful) a master index of provider agencies contained within each I&R's directory

Specify a Standard Third-Party Software Gateway Protocol to Access an I&R Gateway

- The Gateway Protocol is used to negotiate access to resource data controlled by a Gateway. The Gateway Protocol is the only way third-party software can communicate with the Gateway. Initially, just the general requirements need to be specified, not a technical specification.

- Possible capabilities of a Gateway, via the Gateway Protocol:
  - Declare licensing of resource data, if not public domain. If the specific data requested is public domain, transparently pass the resource seeker to that specific section of data.
  - Accept payment from third-party application/case management system for data access;
  - Deliver an access key once license/payment conditions are met. Access data for a defined period of time, or a particular quantity of data;
  - Once access is granted by the Gateway, provide a convenience API. The API will be used by the requesting third-party app to search for and retrieve the desired resource data.

- Once the requirements are fully specified, the protocol can be defined in a technical document.

- The technical document would be used by I&R system providers to encode compliant Gateway functionality into their system.

Provide a Resource Application Programming Interface (API) for Accessing Resource Data

- AIRS should convene experts from the I&R industry and academia for the creation of an API protocol (not code) that permits access to resource data.
  - A resource API is what actually moves the agency/site/service data from point to point.

- Each actual information and referral system would then have to implement the API. The API can evolve over time and have clear versions to ensure both sides are speaking the same version of the protocol.
- At first, the API should accommodate sending and retrieving data simply in an AIRS Schema equivalent format.
- Over time, a Linked Data representation of the resource information should be added to the API.
- The existing software vendors prefer a specific type of payload: JSON, since it requires less memory to process than XML. Both Linked Data and unlinked data accommodate the JSON format.
- The API needs to be freely licensed without restrictions, so all systems can use and extend it as needed.
- The API payloads (aka serializations) should allow the conveyance of custom, or multiple taxonomies, especially for specialized I&R purposes.

- Existing candidate APIs already exist for AIRS to pick and choose from. The following APIs should be examined for useful features or even partial/whole adoption by AIRS.
  - Los Angeles County 2-1-1 has a working resource API. It should be publicly available soon, and is expected to be freely licensed
  - iCarol has a working resource API. It is copyrighted, and not free. iCarol has said it would consider making it available for reuse as an information and referral standard.
  - OHANA has a freely licensed resource API: http://ohana-api-demo.herokuapp.com/api/docs . It does not closely follow the AIRS agency/site/service model, but could either be modified or mapped to the AIRS model. Open Referral is using this API.
  - The Open Community Information Sharing System:
  - https://code.google.com/p/openciss/wiki/openCISS_API_v2 Version 2 has not been implemented, but it follows the AIRS model. Version 1 has been implemented on the server side and on one open source mobile client app: http://homelesshelper.us *Disclosure: the author has worked directly on this API.

Encourage I&R Systems to Publish Resource Data as Linked Data

- Linking Data makes the data navigable and predictably consumable as a resource by the world (not just I&R systems). Linked Data is also more expressive, as it is described in terms of a vocabulary, which can derive from AIRS' existing Agency/Site/Service model.

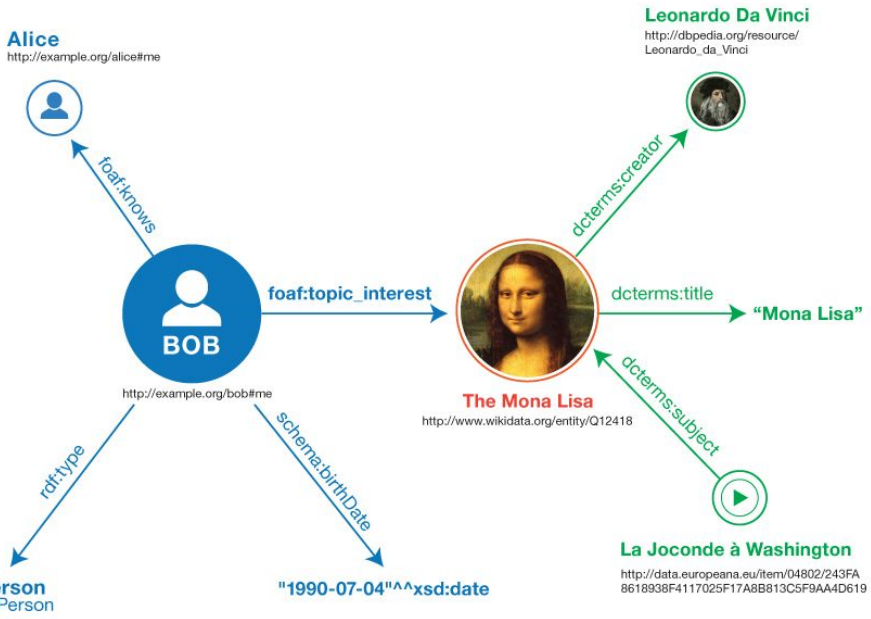
- This does not necessarily mean providing "Linked Open Data", which is Linked Data accessible without any restrictions. Access to the data can be optionally mediated by the Gateway mentioned above.

- I&R systems can continue storing data the way they currently do, but using Linked Data may require providing the data in a different format (RDF), and storing additional relationship information appended to the existing data.

- See the section on "Why Linked Data?" below.

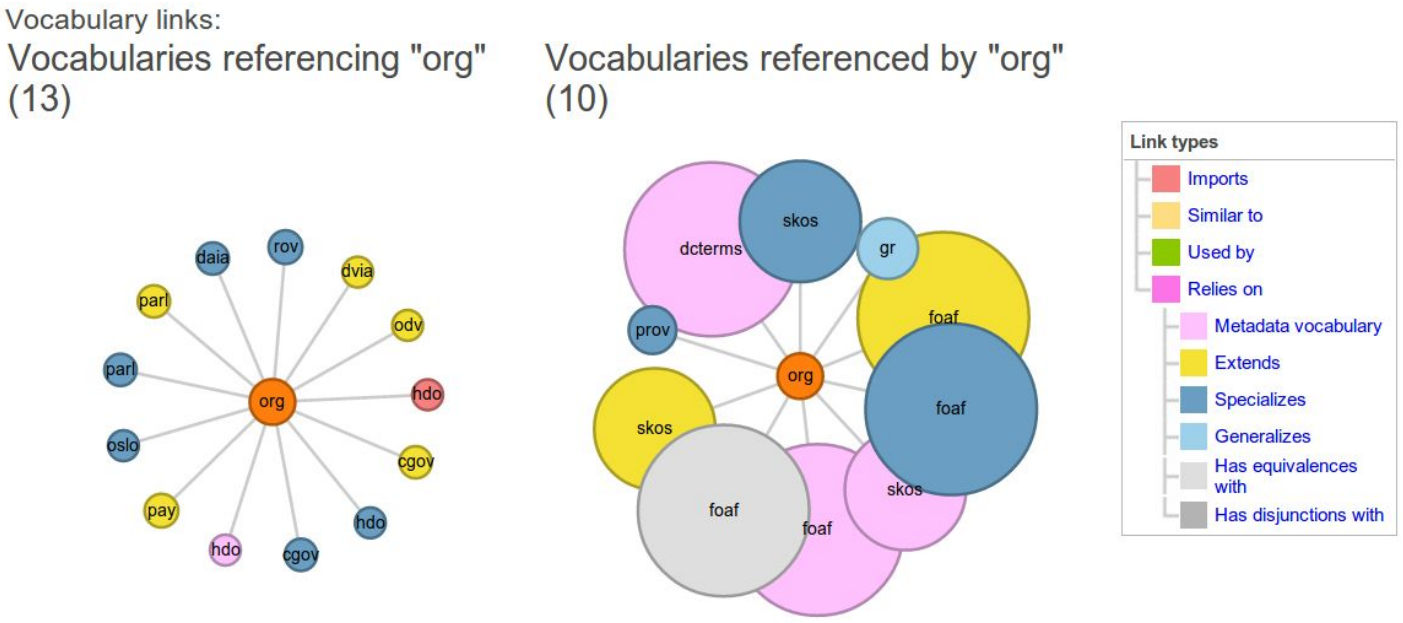
Coordinate the Publishing of an AIRS Vocabulary

- An AIRS specific vocabulary in the style catalogued and diagrammed at http://lov.okfn.org would provide structure to the published Linked Data. It would be an outgrowth of the existing logical structure behind the AIRS XML Schema, version 3.1 (Agency/Site/Service), and can be related to existing vocabularies like geospatial (to define service regions), and friend-of-a-friend (to define relationships between people), vcard (contact info), etc..



"friend of a friend" from http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225

The following image shows how the concept of an organization is related to a series of other established vocabularies:



from http://lov.okfn.org/dataset/lov/details/vocabulary_org.html

The task would consist of AIRS determining what additional vocabulary terms it needs to add, that don't already exist, and make a separate vocabulary for them. AIR's new vocabulary would probably eventually be added to this same web catalog of vocabularies, for the world to reuse.

- Planned vocabularies, such as the budding National Information Exchange Model (NIEM) vocabulary, could be integrated as needed to provide deeper linkages to the mainstream Health and Human Services delivery system at all levels of government in the United States.

- The AIRS vocabulary can evolve over time, become more expressive, and linked in new ways to existing vocabularies. These new relationships could promote a myriad of new uses of information and referral data. For example, allowing analytic software to find formerly difficult to obtain results like, a geographic map showing the national distribution of homeless shelters by building age, and other very specific queries involving normally unrelated data sets. Published Linked Data would be browseable across I&R system boundaries, yet still protected as needed by gateways. This would allow an authorized on-the-fly mobile app to get a list of "all places that provide bus tickets in a 30 mile radius of current location", spanning I&R boundaries and systems. The mobile app would interact with each relevant gateway to obtain the information it needs.

Implementation Phasing Options for AIRS' Recommended Work

The strategic path for AIRS to implement the Directory, Gateways, and Linked Data is not linear, and each part can mostly be approached independently. However, publishing Linked Data generally will require the Gateway, but not the Directory. Also, the Gateway can work without Linked Data (it can just control access to existing data) or the Directory. The Directory is largely independent of the other two components.

The Directory can be a single implementation or multiple separate directories. It could involve adding certain data elements such as gateway endpoint URLS, geographic coverage data, and an universal unique index to sit alongside the existing data set at 211.org. It is recommended the Directory be published as Linked Data (not HTML, which is designed for human viewing). To update the Linked Data set, it may be possible to use the existing web interface already used to populate 211.org, so that each I&R can update their own information.

The Gateway is a protocol, so a workgroup must first design the protocol. Then, each I&R must implement its own. Early adopters can publish how-to guides and specifications, for later adopters.

Publishing data in Linked Data format behind a Gateway can often involve leveraging an I&R system's existing non-linked data set. This could simply require an on-the-fly conversion upon demand (once the Gateway criteria have been satisfied by a third-party application), or it could require synchronization between a Linked Data "Triplestore" and the database, or it could even involve a standalone set of Linked Data set built from the ground up. A convenience API for accessing the Linked Data set should also be made to access the published data.

In this section, some implementation scenarios are described. They are not comprehensive nor mutually exclusive scenarios, but they attempt to illustrate some potential paths forward.

Scenario 1 - Closely Managed Process with Funded Activities

This first scenario presents the most rapid and resource intensive path. Dedicated funds from grant proposals and existing institutional resources are pooled to coordinate the development of the three components. The Directory would be built and hosted alongside the existing 211.org, or another location supporting comprehensive I&R. The Gateway Protocol would be designed by a team of I&R representatives from industry practitioners. The team would also build and test a reference Gateway, meeting the specifications of the Gateway Protocol. A pilot set of Linked Data of community resources would be published for live public access, sitting behind the Reference Gateway.

Lessons learned from these developments would be shared with all practitioners, and competitive alternatives would be welcomed to spur innovation. The three components can all be built from common, available technologies and existing software, so much of the effort should focus on mimicking best practices from other industries.

Scenario 2 - Crowdsourced, Incremental Development

This second scenario involves looser coordination of efforts, with AIRS and other I&R coordinating bodies encouraging the development of missing pieces, prioritizing lagging components as they are identified. Modern tactics, such as calls for crowdsourced contribution, "Kickstarter"-like crowdfunding campaigns, hackathons, project solicitations with software focused groups like Code for America, Hack for Change, and Open Source for America, etc. can be employed to spur momentum as needed. AIRS still would serve as a coordinating body for this system transformation, but would perform more of a shepherding and collaborative role, than the more directive, coordinating role in Scenario 1.

Scenario 3 - Gradual Voluntary Buildout

This last scenario is the slowest and most "organic" approach to achieving the three specified component goals. AIRS and other I&R coordinating bodies would "cheerlead" and recommend efforts moving in the directions described in this and related documents. Low hanging fruit would be picked whenever resources and opportunities present themselves, but largely, the initiative would be driven by competitive interests of existing and emergent information and referral software providers. The Gateway Protocol specification and the AIRS Vocabulary may be the two areas where more intensive involvement by I&R coordinating bodies such as AIRS are required to achieve meaningful harmonization. The Directory, Gateway implementation, and Linked Data publishing activities can largely be completed as isolated activities by actors working under various auspices.

Why Linked Data?

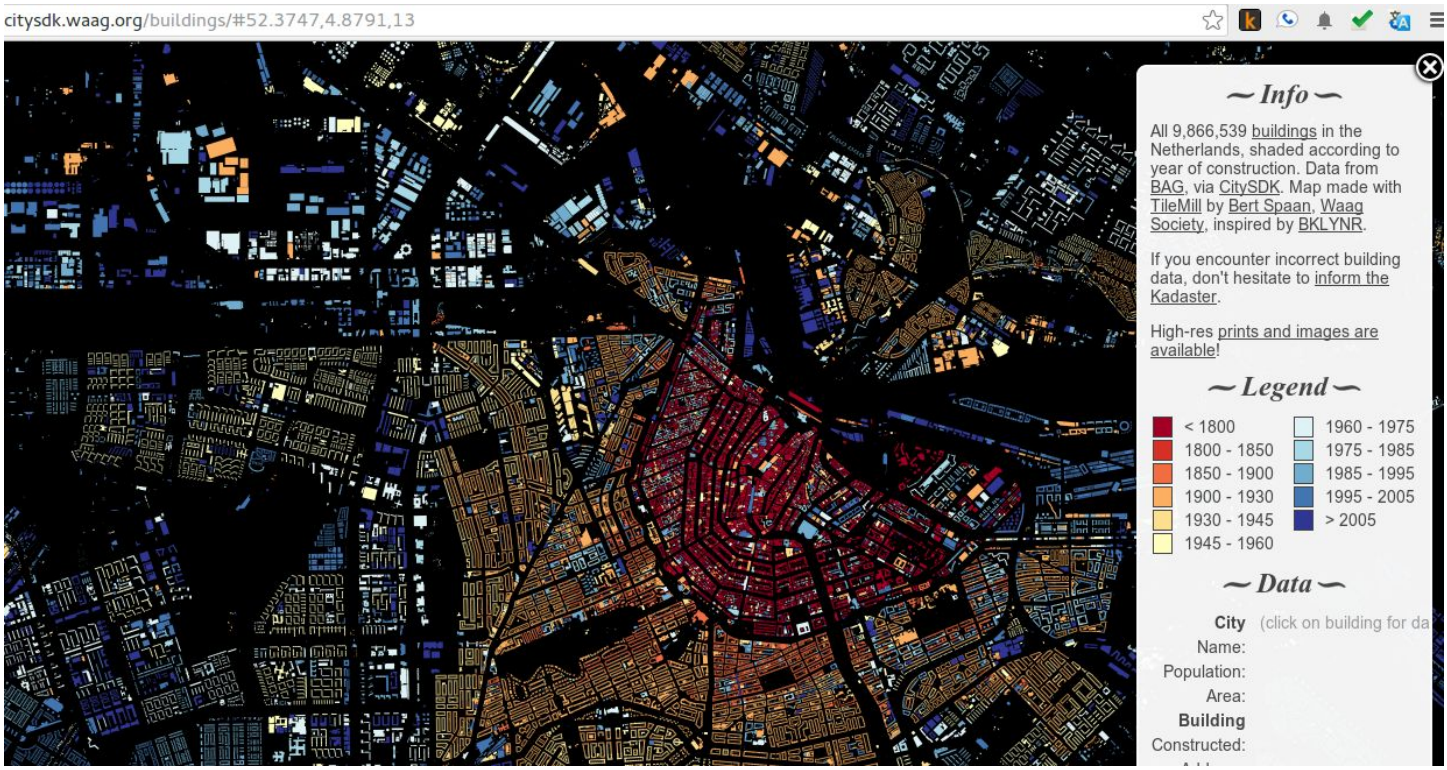
Big Data systems have become increasingly powerful and ubiquitous in recent years, but how does the data stored within one of these systems relate to data within another system? Advocated by Tim Berners-Lee, creator of the World Wide Web, Linked Data can turn the World Wide Web into a virtual glue, cobbling together these massive data sets, publishing their contents to the outside world in an accessible, expressive manner. Linked Data is essentially data with meaning and having a defined place within an established vocabulary. Linked Data also points to related data, so more can be known about the subject.

For example, IBM's Watson uses Linked Data to come up with recipes, by looking at available similar recipes, looking for patterns in the ingredients, and following related concepts. New versions of Watson will enable it to handle more random questions on the fly.

> "Some of what we're going to be doing is taking government data that is now available around the world and making it a little more machine-readable with this Semantic Web stuff. You'll be able to ask Watson 'Where can I find data about obesity in Europe?'"[2]

The idea with Linked Data is to take big sets of data, and for each piece of data, describe it further by linking it to online definitions (vocabulary), and to other piece of data. To glue the first piece of data to the second one, you use a "predicate" which contains a verb. An example could be Bob [subject] is a Dog [predicate]. Bob is the first piece of data, the predicate is the rest of the sentence, which contains the second piece of data, Dog. Dog is the the object, within the predicate.

Practically speaking, Linked Data has shown itself very useful for exactly the kinds of problems AIRS confronts. Large corporations such as Wal-Mart have been using Linked Data to coordinate supply chains [3] and healthdata.gov is a large compilation of Linked Data, among other things. The Netherlands have made their entire building registry available as public linked data. The registry data set consists of about 573 million triples (triples are the subject/predicate/object we just discussed). The registry's building age and geospatial data set have been used to produce the visualizations shown below.



Age of Dutch Buildings, visualization of the Linked Data set located at http://lod.geodan.nl/basisreg/bag

[2] taken from http://semanticweb.com/ibms-watson-the-smartest-thing-on-earth_b38768#more-38768
[3] taken from http://sandhill.com/article/why-the-titans-of-business-wal-mart-to-dow-jones-are-investing-billions-in-semantic-technology

screenshot from http://citysdk.waag.org/buildings/#52.3747,4.8791,13

Similar analysis could be done using linked I&R data to show a "heat map" of service usage for a given AIRS Taxonomy code across multiple I&R implementations.

How Do We Provide Linked Data?

"Linked Data" is one of those terms that sounds like a fancy name for something we already know as hyperlinks and URLs. Linked Data is like that, but it does require a few specific things to work, that most people don't associate with simple data linkages.

To get started providing Linked Data, you need to define your "working lingo" in a linked vocabulary. Right now, AIRS' equivalent to this is the AIRS Glossary of Terms and its technical representation, the AIRS XML Schema. But, those two constructs aren't technically linked to other terms and concepts on the web. To link them to other concepts, a linked vocabulary needs to be published. One popular way to do this is to publish an OWL file, which declares your vocabulary.

Then, you publish your Linked Data by creating a location on the web that points to where you keep that data. Then, make sure the data is in a format that describes relationships between things. RDF is the most common, but there are equivalent formats for every software developer's taste. Software can read these formats, discern the relationships and follow the links to the related items. RDF can be sent on demand, or it can be embedded in a web page.

Related Concepts

Linked Data versus Search Engines

Linked Data and search engines can work together, and don't preclude each other in any way, but Linked Data expresses the data according the agreed upon set of interconnected vocabularies, whereas search engines typically want the data tagged in their suggested way (one popular way is schema.org, a lightweight vocabulary, more like a taxonomy than a linked vocabulary or ontology). Linked Data doesn't rely on a search company to crawl the data beforehand, since Linked Data can be directly searched: it's already indexed and related to similar concepts. This feature of Linked Data gives local information and referral providers some measure of autonomy and control over their data.

Summary

With Linked Data, AIRS has a tremendous opportunity to enhance the accessibility and reuse of its data. The underlying linked data vocabulary AIRS could establish would enhance the clarity and specificity of the data, making reuse in other contexts much easier. Linked Data would also encourage a federation of local information and referral data aggregators, acting with the same fluidity and effectiveness as a single unified data warehouse of crawled referral data. The Gateway would protect the original terms of the data reuse, while providing access to automated third-party software tools. Third-party software tools would use the Gateway Protocol to negotiate terms of data use with each Gateway. The Directory of information and referral services would provide a convenient top-level reference for locating the various

I&R gateway locations, and provide some general summary information to help software search tools to determine where to request data.