terradue 2.0

www.terradue.com

On Spatial Data Search

Jo Walsh

This essay takes a closer look at the apparent lack of cost effectiveness of the Spatial Data Infrastructures effort in developing "catalog services" and community specification for spatial search.

While conventional text search engines move to assume the same functions the GIS industry should reach out to other information retrieval and Internet communities.

Short Title	White Paper Template
Prepared by	Jo Walsh
Approved by	Pedro Gonçalves
Reference	T2-Research-07-003-OnSearch
Issue	1
Revision	0
Date of issue	2007-09-07
Status	Final
Document type	White Paper
Distribution	Public
Category	Research
Keywords	metadata, geographic information
Comments	The third document on the Duetopia Series about Geospatial Services and Applications

Table of Contents

Introduction

Note: *This is the last of the three essays, "On Metadata" and "On the Development of Geospatial Ontologies". They are a set, and though a reading of the first two is not necessary to enjoy this one, the conclusions reached there constrain what is discussed here.*

One joins in a spatial data infrastructure in order to share data, and to find it. Our own use of others' resources increases a network effect, encourages us to contribute our own data for the use of the network, see it built upon and augmented.

What data is available to "discover" on a "geospatial web" right now? Some well-known data sets are referred to in many places, available from many sources. For core information with a common purpose - in the public domain and at low resolution we think of VMap0, Landsat, SRTM - many copies are distributed. Other data sets in low or local demand may only be found in a few places.

In the discourse of Spatial Data Infrastructures, we have "catalog services" - directories of information, pointing to where spatial data and services can be found. For metadata "records" describing spatial data and services, there are "registries". The GIS industry coins specifications for search interfaces, where it might do better to reach out to other information retrieval and Internet communities. A bigger picture of data search surrounds the spatial problem; and conventional text search engines move to assume the same functions.

We try our best - but why?

Reflect on the early days of the web as a medium for sharing information. Directory services, registers of links on topics, "site of the day" sites. When information sources numbered in the tens or hundreds of thousands, this was sustainable. There was no distinguishing between spatial or topical scope of pages, their source languages, the inclination of participants. Full-text search, link crawling and cross-referencing and indexing engines, ate the web and became most users' primary interfaces to it; the key point of introduction to a new medium.

Public web spatial data services are "discovered" by typing "inurl:REQUEST=GetCapabilities" into Google, or by looking through lists of links on HTML pages. A layer of "Service discovery services" is offered, threatening infinite recursion, and no resolution to immediate problems of finding data or of making it known.

Meanwhile, the work of Google employees on map-based search of KML and GeoRSS expands, both these specifications now being collectively maintained by the Open Geospatial Consortium (OGC). If these standards and methods are working, why not pursue them? One may not wish to rely on one large, off-site provider for access to the contents of one 's own archive. The option of releasing data onto the public Internet is not available to many agencies.

At this point, it seems beneficial to step back a bit, look at the bases of spatial search, and consider grounds for a different kind of solution. Grounds on which we can build something that no large and distant collector and indexer of data can do a better job of - local knowledge in the systems within which we all work.

To design better search facilities for spatial data, look at the goals which an engine, minimally, needs to meet:

1. Connecting information to those who have a use for it, whether the use case is beknownst to them or not
2. Filtering the potential collection of stuff according to what the requestor is more, or less, likely to want back.

"Publish, find, bind" is a phrase used to describe how web services should work, in information retrieval circles. The viewpoint of "publish, find, bind" is of the service, the perspective of the programmer. Stefan Keller identified a cycle of "registration, declaration and citation" of geodata in a data harvesting and "discovery service" context, a description of the cycle closer to usage patterns of services. [OSGeo-MD]. Often the discussion of metadata, domain modelling and discovery is biased to the toolmaker's, not the tool-user's view. How to bring the two closer together? A design aim is to stay as close to the data as possible at all times. If a new kind of spatial search engine needs a different kind of client to talk to it, what can we see in existing and near-cutting edge GIS client applications?

Recommendation and its discontents

"Social Software" formed, along with "Web 2.0" and the "Attention Economy", a hype bubble that occupied the fallow space between the first two dot-com booms. Services that connected a "social network" to collective annotations - del.icio.us for web resources, flickr.com for images - were duly eaten by the large search engine companies. The ability to connect a person to a citation can have profound consequences. Yet the potential of these services to filter, as well as generate, data collections has never been proven. What other hidden utility do they hold?

- At what point does data use become a *de facto* citation of it, or repeated use a recommendation of it?
- What kind of context on either side do we have available at the point or interaction / of service usage?
- How many hoops do we have to jump through in order to establish a useful model mapping between the terms the client wants to traffic in and the terms the service provides?

In their White paper on User Evaluation of geodata, the Geodata Commons group eschew the formalism of a user recommendation or rating system for geodata sets or packages [CGD-Eval]. The proposal includes some freetext evaluation of how useful the data set turned out in a particular context. Given a large enough volume of participants enough common context might develop for this approach to be useful.

Our purposes in handling data vary greatly, different participants' depth knowledge vary, and "rating", even for a "particular purpose" seems unhelpful. To make rating-based recommendation useful for filtering a large volume of data, rating of raters is needed. It is not clear that any of this is helpful in building better data access and distribution systems.

Suggestion based on popular rating tends to induce power law behaviours where they need not apply - the majority of users see the minority of data sets in their results; the system sees self-reinforcing list of high priorities. The best kind of "Recommendation" does not need to be stated but can be inferred from a real commitment of resources - two books bought together on Amazon, two papers cited together in an academic journal.

What's my provenance?

Attribution of information found on the internet is more useful to a new model search engine. Knowledge about how data is being reused is critical to its future use. A data set's provenance is where it came from, attribution for its creators. The lineage is a history of other data sets and processes which have been used to compile a "new" data package from existing parts. The provenance and lineage of data is important to anyone engaged with a data search engine.

On the wider Internet, search providers do not often pick up attribution data for the content they are collecting - even when it is possible to clearly provide it in a standard way [Platial]. In transmitting data between sites, much of what is known about the data at the source is not transmitted. Yet waste for one is wealth for another, and "Drive-by metadata" may be fruitful.

For packages of data, a record of their lineage is a record of the cooperation, reuse and edit history of different parts of compiled data. As data passes through applications, an implicit audit trail is created. Other applications and services see part of this trail, not all of it. Even in an interaction as simple as a request for a set of images of a collection of layers from a WMS, useful context about how others use information is created.

Thus a search engine should be able to see the transaction history behind spatial data, while it is being created and annotated, while it is being published. Personal provenance information should always be collected with data sources - ideally not be separable from them. Clients software working with spatial data should be able to "playback" sessions of compiling different sources and applying different processes, contributing session history to a shared repository.

Against classification

Each act of classification is an investment of semantic energy. The stored semantic energy gets used up whenever one actor presents a classified resource to another and asks - "is this appropriate for you?" Sadly we cannot all be Wikipedia. to have a well of semantic energy to draw on, seemingly without limit, from a network of hundreds of thousands of contributors - though we can all aspire to this state and to reach it with more efficiency, with less wasted effort on the part of those participating in common description.

To what extent is a lot of data “catalog” maintenance appropriate, either to its maintainer and source publisher, or for a client to whomthe data is potentially of use?

- What is the cost/benefit to the originating organisation?
- What is an alternative, or synergistic supplement, to taking a catalog approach to finding geospatial data?

There are risks inherent in relying on indexing and searching the text accompaniment to a spatial data source. We don't exist in data poverty, quite the opposite, but we feel it because the techniques developed for effective text search don't apply to us so well in geospatial. We solve simple problems, glossing over larger ones. Text search of accompanying metadata lets us pick the low-hanging fruit, but we should taste the fruit first before picking more.

A core consideration for a repository or archive maintainer is ease of movement - between information models, interfaces, standards. Update to a service needs to be non-breaking and non-work-rotting. Thus the "stack" approach that has arisen in open source software, components with different specialisms that grow organically together at the edges, facilitated where possible by open standards.

Exchange descriptions of data, or data?

To convey more domain knowledge, more context to a search service is to get better knowledge back. An exchange of data, direct, is the starting point in the search process. A data model provides at least a basis on which to state "more like this", and filter a long list of potential results.

If the client has no data or domain model to start out with, it should be straightforward to pick a starting point from the service - a sample package, which one may or may not be free to inspect the contents of (a generalised or spatially limited subset of a larger, popular data set). This can be seeded on the basis of a query for keyword, spatial/temporal extents, or picked on a random basis from a pool which may be limited by other criteria. [Similarity]

The passage of one client through this process leaves behind trails for others. For the developer of a search/discovery engine, it offers rich "loose inference" prospects, linking data sets together for subsequent users starting near the same place.

In this kind of "model-exchange query", it is essential is that both sides of the interaction learn as much as possible from it; not simply the client which comes to a search service and takes something away; nor the search service which gains knowledge about "relevance" from a client's interactions, and does not offer this knowledge up. The receiver of data, through learning more, is able to provide a better kind of service. Observation of usage patterns helps conventional text based search engines to provide more 'relevant' listings on a bulk basis - links that are "clicked through" will get higher subsequent "relevance" ratings for later searches on the same topic.

The distribution part of a data search service, given the ability to collect data sets together on the basis of tacit user suggestion, can merge collections into:

- Packages of data and configuration / metadata [MEF]
- Graphs or series of links to be traversed
- Suggested resolutions in apparent cases of conflict between data sets

If one has analysis results derived from a dataset, it should be straightforward to contribute them back to such a service. Data available before a "validation" or classification phase has been carried out, can be made available to a limited constituency, become subject to semi-public correction.

What kind of service sits on the other end of this act of contribution that is simultaneously an act of searching?

Open annotation

One hopes for spatial search to be put to a lot of unexpected uses. In the design of an effective system we have current re-use cases, from the providers of geographic information and the analysts and agents who are working with it. Many of our problems are generic; yet generic solutions cannot be imposed.

One way to learn about data modelling practise is to establish "open annotation" of a reference data set, which may be provided by an agency or developed in common by a data usage constituency. The success of public participation GIS projects such as OpenStreetmap encourages more data holding institutions to ask: "Who has enough vested interest to annotate and correct our data?" and "How can the effort be distributed amongst all those who have a partial interest?".

"Open annotation" is a practise growing in currency in data-heavy research. One shining example is that of the Human Genome Analysis project. In an atmosphere of "fierce collaborative competition", agencies with a mixture of public and private funding compete to analyse chunks of a public domain base data set.

Tim Hubbard, the coordinator of the Cambridge analysis centre, has spoken of the manifold increase in research speed as a result of a network competing to do analysis and to aggregate results. He used the analogy of a "coordinate reference system" and a base map, to describe the layer of public domain information available to members of the project. Participants submit contributions and keep track off others' via a simple RESTful interface, designed to be distributed around nodes and collated, or aggregated, on a regular schedule. [Hubbard]

In the domain of genomics analysis, all participants share a clear common domain model. Yet the principle can be used by any research network which can establish shared policy and agreement on a set of shared models. For GIS systems able to expose themselves to the internet, the active principle can be fruitfully applied to all kinds of spatial data - the collaborative verification of land cover classification, translating the semantics of features, amending their geometry, or the patching together of contributed higher-resolution tiles into a seamless image.

If there is to be an "SDI 2.0", then this is one direction in which it will go [Fonseca]. The first web client was not a "browser", it was an editor. The first "search" client on the web of data may be a contribution tool more than a consumption tool. A data user's interest is reflected in the contributions they have made. A data search service would benefit from intimate connection with a data annotation service.

Complex models, limited protocols

"Search" still makes one think of the web; current practise in spatial data search is too dominated by web search interfaces and metaphors. The solution of one massive index, containing everything the searcher is likely to want to find, has seemed a good working solution for finding distributed resources on a big network. What results is a lot of web-facing "geoportals", offering a keyword search over abstracts. The more apt interfaces offer a filtered search through a shape drawn on a map, or a span of time.

If source data is not available, at this point the metadata provided by the author is supposed alone to be enough to "evaluate" whether or not the data is fit for use. If the data package is openly available, it is downloaded via the web, saved to a filesystem, opened in a GIS data client. "Web services" offer a faster round-trip for data, more "seamlessness" for the client; but local data and the ability to work off the network is critical for many. The "richness" of the data is lost through the thin view of the web interface.

This problem of complex models, but limited protocols for querying them or carrying them around, is one thing that has held back the development of geodata "catalog services". It is a problem is that data publishing is not more directly connected to data production tools. Metadata is put in the hands of archiving specialists, and everything looks the same. Spatial queries and filters are add-ons to a standard "search" process.

Implicit metadata

Information about how data is used is picked up in the background, transparently to the user. Given a means to persistently and uniquely identify data sets across sites, very useful information about what data sets are related to each other, are usefully combined with one another, which are in frequent use by a person or organisation. Knowing this helps us cluster and filter data sets within suggestions from a search service. Usage patterns help mark out trails for others to follow. M. Gould et al pursue work in this direction under the heading of "implicit geo-metadata". A memorable phrase is "the metadata grows as the data is used", and their work includes plugin development support for open source GIS client software.

The GeoDialogue approach seems like one to watch. It emphasises context of spatial queries over core data semantics and domain models. One can take a wider view of context for machine interactions - what actually happens in client software during analysis acts, filtering and combining of features, and the portrayal of them. If implicit context during a transaction is being monitored and recorded - where is it being stored, and how is it being shared?

State and aggregation

In applications that have been built on RSS and to some extent GeoRSS one sees semi-structured data being created and stored in a distributed way. "Aggregation" of feeds leads to services like Bloglines, keyword oriented pop contests like Technorati, and local group newsreading clients. These are all doing essentially the same thing - using some perception of the user base's intelligence to improve the service and help them support one another. What differs is what is sent back to a central source - knowledge about who is reading what, about what different topics are grouped together - and how that information is being exposed to, and reused by, the client.

Search services as they are apparent now are collectors - byproducts of the linkedness of the web. They traverse links between chunks of data, and can load up their understanding of those links with optional extra values. As chunks of geographic data don't tend to refer to one another, one cannot take directly a "crawler" approach to building indexes of lists of data sources. In the current "standard" model, files and web services must be registered somewhere, with an index of their properties, to be picked up by others.

This does not look so dissimilar from OpenSearch, where each node providing an interface is running a registry of their resources, to be "harvested" by A9 and other large search providers / metadata brokers. This model tends towards the development of more centralised "one-stops", competing to collect and store the most data in one system.

As for the providers of spatial data distributions, potential spatial data search and discovery hosts; are they resigned to this? It is possible through an assortment of techniques designed to "localise" the use and transfer of data in different ways, The behaviours of a network are not available to just one node in it.

To what extent need we rely on broker services - for source data and for metadata? If policy insists the two be separate, then we rely more on central data brokers. If data is going to be annotated, collected, amended, then it makes sense for all those amendments to go back to the source of the data. Can we rely on large aggregators to send user corrections and contributions back to the mapping agencies from which the data comes? Administrators at such agencies have been reluctant to release data free for re-use and redistribution, on the grounds that commercial middlemen can work with the original data, extend it and "add value" to it, sell services, and suffer no obligation to return fixes to the source.

Yet they miss out on a practise that can massively improve data currency and accuracy at minimal investment, and miss out on the opportunity to build, or participate in the building of more reactive services themselves. Running access services nearer the source, encouraging contribution as a path to search suggestion, in simple ways agencies can recover more of the 'added value' to their data for themselves and their registered users.

But what about my rights?

Users of spatial data need assurance about their rights to reuse data. Providers who claim the right to limit who does what with the data they provide, need a means to assert those limits. In current data licensing practise those rights limitations can be expressed using a URL, and thus be interpretable and reusable by software. Data attribution and lineage information is carried around in feeds, stored alongside it in repositories.

However, seeking to limit access to data based on rights attached to data directly, introduces layers of extra negotiation. Payment, the exchange of contracts to establish a "Digital Rights Management" (DRM) mechanism - these act against the ability of search engines and repositories to index and mirror data.

This research thread has been about different ways to lower the abstraction barrier between source data and the way it is represented, summarised. To realistically assess "suitability for purpose", access to the source data is needed. Given a large broker, one could look at a "Google Print" type solution for a "GeoPortal" system - one trusted third party is given the right to access and index a large body of otherwise "protected" work and offer public excerpts from it. The provision of limited subsets of datasets - either a small subset in space, or a "generalised", less precise version of the whole - under an open access license would be a useful general practise for better "evaluation". [CGD-Challenges]

Yet one wants to distribute the ability to "add value" to spatial data to as many "trusted third parties" as possible. Particularly so, given the benefits in developing small-group, very data-dense search services. There are alternatives to perdataset licensing and pricing, given the right networking agreement. Since the rights to reuse and redistribute a data set are tied to a context of an organisation on a network (usually the Internet), consider using the place where sharing happens as a different way of "pricing" access. Purchasing a flat-rate subscription to a network, one can gain access and redistribution rights within that network. This is a model that is emerging from media distribution, but that could map easily onto existing municipal, regional and inter-agency data sharing agreements. [Playlouder]

In such an effort it is important to work closely with the constituency of a data set or sets - those people producing, correcting and working with the data most closely. Their desire to share more fully with a network of collaborators than with the wider world can be simply handled. "Privacy" can be maintained in generalised or restricted subsets, rather than incurring a blanket exemption on the non-privacy-invading parts of a complete dataset.

In a spatial data distribution system, archiving is going on all the time. Tile distribution, caching of vectors, emission and merging of updates, and everybody has a store. Catalog services, text string query and model based query interfaces all expose indexes. All users of a system have capabilities to get at them attached to their credentials on connecting to a network. Chains of capability connect directly to chains of accessibility, vouching for others to access repositories or processing capacities. Passing a set of instructions down a chain, transparent to a user once credentials are initially established.

A view from the space station

More data search becomes spatialised. The spatial distribution of network nodes comes to affect more, the workings of a service. Spatial indexing and query techniques come to apply to a lot of other kinds of data sources, which may be only partially, or not at all, spatial. The techniques built in the GIS domain for partitioning a search space out into parts and see if what one wants is in a bit of it, reapply. Algorithms for assessing similarity and proximity of features can be applied against all kinds of research, experimental and open data.

Thus the view from the space station; how best to get off the ground? Recall these two core aims of a spatial data search service:

- To make useful data easier to find
- To make irrelevant and distracting data easier to filter away

We cannot rely on others to expend semantic energy - in registering spatial data sources, describing them and citing them - overtly. To what extent it is possible, such context must be quietly captured and contributed via the client application. Sets of common requests correspond to generative packages, rather than to any idea of recommendation on a per-individual basis.

If discovery is the better part of access, then distribution is the better part of discovery. Through reverse engineering data access we can provide means to better discovery. If source data is openly available, it can be archived where it is frequently requested, lessening the burden and the wait on a distributed search.

Static services, on a static network, are not what we expect to find. This research arose from a project to distribute land cover information via BitTorrent, breaking up the data set into spatial extends of interest. BitTorrent is data neutral, and as well as a literal data transport it serves as a metaphor - for archives, and queries, being collected locally within peer groups according to shared interests. More than anything a data collection, index and interface provides a shared focus amongst its participants.

The approach is not concrete, and nothing can be called a conclusion until it is proven or disproven by software. In combining the different tactics above, the results will be like dipping a ladle in a data stew; depending on the taste of the results, ladling more of the same or different on the next try.

References

Writing and Presentations

- [Platial] http://platial.typepad.com/news/2007/04/best_practices_.html "Best Practices for Emitting KML and Giving Props to the Geoweb's Authors", Chris Goad, April 13 2007
- [OSGeo-MD] http://wiki.osgeo.org/index.php/Geodata_Metadata_Requirements, Stefan Keller, 10th Oct 2006
- [CGD-Eval] http://geodatacommons.umaine.edu/wpapers/CGD%20User%20Evaluation.pdf Commons of Geographic Data White Paper on User Evaluation,
- [CGD-Research] http://geodatacommons.umaine.edu/wpapers/CGD%20Research%20Challenges.pdf Commons of Geographic Data Research Challenges
- [Fonseca] "Is it time for an SDI 2.0?", paper presentation at 13th EC GI-GIS workshop, 5th July 2007
- [Gould] "Implicit Geo-Metadata", paper presentation at 13th EC GI-GIS workshop, 5th July 2007
- [GeoDialogue] http://www.geovista.psu.edu/publications/2006/2006-geoinformatica-cai.pdf "Contextualization of Geospatial Database Semantics for Human-GIS Interaction", Cai, 2006 (and previous papers in the GeoDialogue series)
- [Hubbard] http://okfn.org/ok1/after/ Presentation in panel session on "Open Scientific and Civic Information", Open Knowledge 1, March 17th 2007
- [Similarity] http://www.similarity-blog.de/ General paper collection and blog on "semantic similarity"

Online Services

- Google Print - http://books.google.com/
- Technorati - http://www.technorati.com/
- Bloglines - http://www.bloglines.com/