

Active PURLs for Clinical Study Aggregation

David Wood¹ and Tom Plasterer²

¹david@3roundstones.com, ²Tom.Plasterer@astrazeneca.com

3 Round Stones

AstraZeneca

The problem: No coordinated view of clinical study information. Information is distributed across departments, subsidiaries and government data sources.

The solution: Gather, convert, aggregate and format for display

3 Round Stones and AstraZeneca created a system to allow coordinated views of distributed clinical trial information. The system extended the Callimachus Project, an Open Source management system for Linked Data.

Persistent URLs, or PURLs, were used to provide globally unique and resolvable identifiers for each clinical study. The PURL concept was extended to enable PURLs to have multiple targets and for the results of each target to undergo arbitrary transformation. PURLs which have such capabilities are called Active PURLs.

Information sources relevant to clinical studies were identified, regardless of whether their location was internal or external to the pharmaceutical company's network. Active PURLs were used to resolve data sources having HTTP endpoints capable of returning XML or textual results. Each information source is dynamically transformed into Resource Description Framework (RDF) formats and all sources' results then merged into a single, temporary graph of RDF data. Information is rendered to end users as coordinated HTML descriptions regarding each clinical trial using the Callimachus template engine. Machine-readable versions of the data are also available.

How semantic technologies help

Linked Data techniques can help to address both the availability of clinical trial information and provide a means to build effective information systems using it. Linked Data techniques allow for "*cooperation without coordination*". Publishers of data provide context for use by third parties in other portions of a distributed enterprise. Users of Linked Data can combine information from multiple sources. Subsequent publication can create a virtuous circle of positive feedback, allowing researchers, informaticists and support staff to collaboratively and distributively build a reusable knowledge base.

Challenges

Distributed queries have many known limitations, such as the introduction of multiple single points of failure in any given PURL resolution. HTTP timeouts, auth/auth errors or other network failures can slow or stop a pipeline from returning correctly.

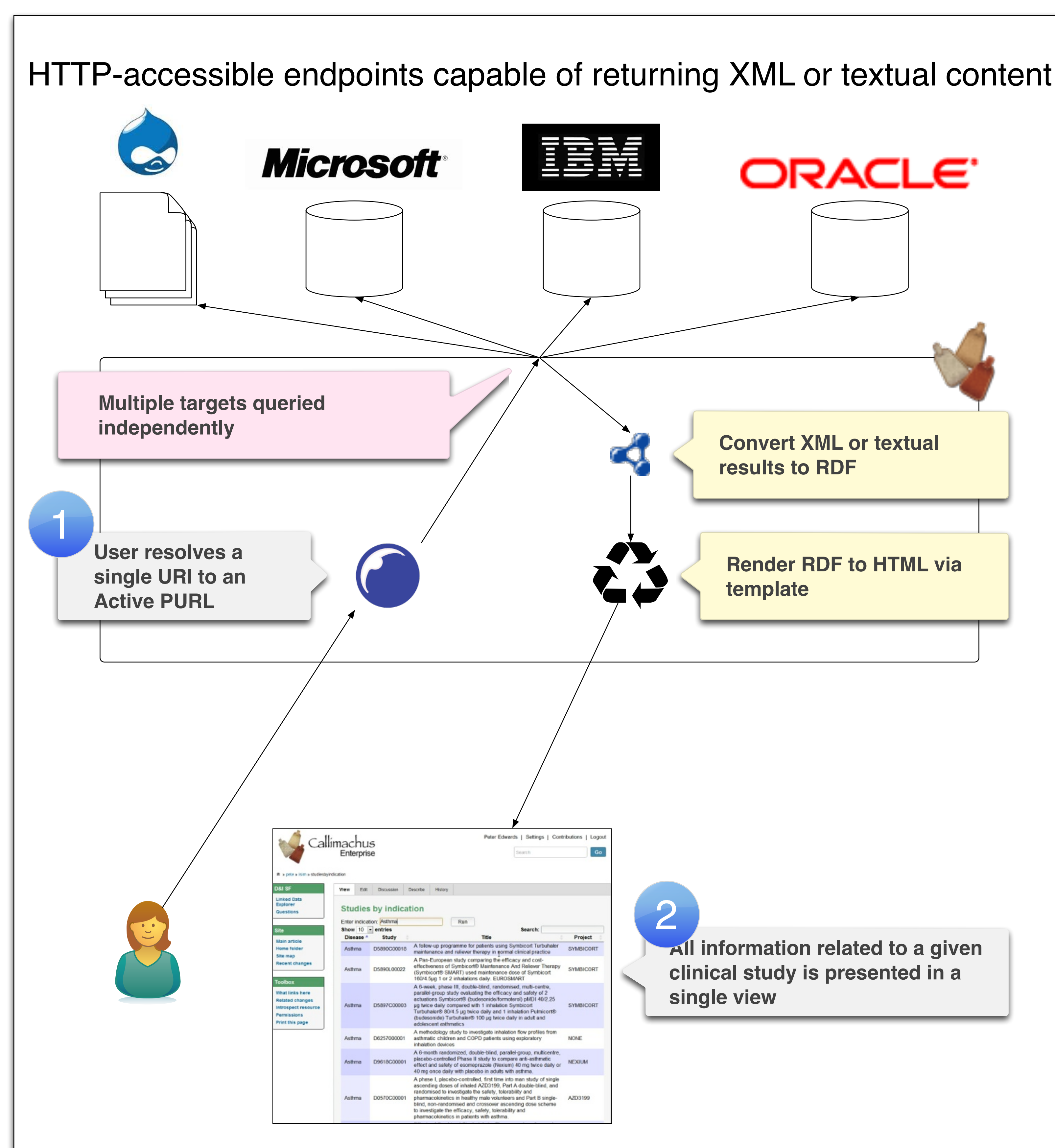
Similarly, distributed queries can result in variant query-time performance due to complex network and endpoint performance variances.

Proactive caching and cache management strategies can improve runtime performance and protect end users from the limitations inherent in a distributed query architecture. Caching of intermediate results from endpoints has not yet been implemented.

User experience

1 Users resolve a URL that provides a unique identifier for a clinical study, drug, chemical or other concept managed by this system. The user may be presented with the URL on HTML pages, search it via full-text techniques or discover it via semantic search.

2 Users are presented with a dynamically generated Web page representing aggregated clinical study information. Users are isolated from the complex and distributed information environment.



Next steps

We intend to continue to address system fragility and performance issues by implementing improved caching, cache management and graceful recovery from HTTP endpoint timeouts or related errors. Most of these issues can be addressed in Callimachus as that project matures.

We plan to continue to search for additional data sources and refine routine system management tasks.

References

1. Callimachus Project, <http://callimachusproject.org>
2. Persistent URLs, <http://purlz.org>
3. National Center for Biomedical Ontology, <http://www.bioontology.org>, <http://purl.bioontology.org>