Downscaling Entity Registries for Ad-Hoc Environments

Philippe Cudré-Mauroux[1], Gianluca Demartini[1], Iliya Enchev[1], Christophe Guéret[2], and Benoit Perroud[3]*

[1] eXascale Infolab, University of Fribourg, Switzerland
{firstname.lastname}@unifr.ch

[2] Vrije Universiteit Amsterdam, the Netherlands
c.d.m.gueret@vu.nl

[3] VeriSign Inc., Fribourg, Switzerland
bperroud@verisign.com

Abstract. Web of Objects and Linked Data applications often assume that connectivity to data repositories and entity resolution services are always available. This may not be a valid assumption in many cases. Indeed, there are about 4.5 billion people in the world who have no or limited Web access. Many data-driven applications may have a critical impact on the life of those people, but are inaccessible to those populations due to the architecture of today's data registries. In this paper, we point out the limitations of current entity registries when deployed in poorly connected or ad-hoc environments. We then sketch new architectures based on IPV6, structured P2P networks and data replication for entity registries that could run in ad-hoc environments with limited Internet connectivity.

1 Introduction

Data registries are critical components of the Web architecture and are widely used in every-day web activities. For example, domain name registries are databases containing registered Internet domain names. They are necessary for all web users wishing to visit a website knowing its URL (e.g., hostname) rather than its IP address. Thanks to the Domain Name System (DNS) infrastructure, such information can be obtained by recursively resolving a domain name to an IP address.

Another example of registry is the Digital Object Architecture (DOA, see Section 2). It allows to assign unique identifiers to digital objects (e.g., scientific publications) which can then be univocally accessed by users. Their identity will thus last in time even if their physical location may change (similarly to the IP address associated to a domain name, which often changes over time).

In situations where data registries are not continuously accessible, the user experience can be strongly limited. As a basic example, if the DNS server used

* Authors are listed in alphabetical order.

by a client computer is not connected to the rest of the DNS hierarchy, then a very restricted set of Internet domains can be resolved to their IP addresses.

In addition to those traditional registries offering hash-table like functionalities, online infrastructures and applications are increasingly turning to more flexible registries containing information about general objects or entities (i.e., *object registries* and *entity registries*) to power data-driven applications. Emerging examples of that trend are DBPedia[4] and Freebase[5] which, given an entity identifier (e.g., a URI), provide semi-structured metadata about the entity. Another example are the registries used for the Web of Objects, which mediate information between networks of digital devices connecting to each other, enabling information publication or integration in sensor networks or smart building contexts for example. In those cases, the infrastructure needed is more complex than a traditional "Hostname-IP" DNS system and is closer to a global registry mapping unique identifiers to arbitrary structured data.

In situations where such registries are not continuously accessible, however, the user experience can be strongly limited. As a basic example, if the DNS server used by a client computer is not connected to the rest of the DNS hierarchy, then a very restricted set of Internet domains can be resolved to their IP addresses. In an object registry context, discontinued access to the registry typically results in the impossibility to publish data or issue object queries.

In this paper, we argue that data registries increasingly represent an essential part of today's Internet ecology (see Section 2 for a few examples), but that their current architecture precludes their use in many important contexts. For example, we can envision ad-hoc environment where the nodes self-organize without having access to third-party registries. In such transient and poorly-connected environments, nodes have a clear need to discover, connect, and exchange (structured) information with related entities locally and should be able to do so without resorting to any outside registry. Another interesting context is data-intensive object applications, where nodes have to discover and exchange data about very large numbers of entities and should be able to do so in a peer-to-peer manner whenever possible.

We propose in this paper different approaches to overcome the limitations of existing registry solutions for poorly connected environment or localized environments. First, we point out the limitations of current registry architectures, before proposing solutions that take into account the limited connectivity of the peers and enable the management of digital information in ad-hoc environments.

The remainder of this paper is structured as follows. In the rest of this section we motivate our work: we show why it is important to consider entity registries for ad-hoc environments and which are their benefits. In Section 2 we briefly describe existing architectures for data and entity registries. Section 3 highlights the problems of current solutions when applied to poorly connected or ad-hoc environments. Section 4 presents two alternative solutions to exploit entity registries based on IPV6 addresses and on P2P networks respectively. Finally, we conclude and discuss future work in Section 5.

[4] `http://dbpedia.org/`
[5] `http://www.freebase.com/`

1.1 Use case: Internet-less Mesh Networks

A rapidly increasing number of applications—such as open social applications, applications relying on governmental data (data.gov) or Linked Open Data[6]—assume an ubiquitous and continuous access to the Internet in order to power data sharing and data-driven applications. As pointed out in our previous work [6], this is not a safe assumption as there is more than half of the world's population who is cut-out from wide area networks. However, people who do not have access to the Internet still generate data and need to consume knowledge. For example, children who benefit from the "One Laptop Per Child" project, which aims at providing low-cost laptops (called XO) to developing countries, can be connected to each other. XO laptops are used at school while connected to school servers, but also at home where connectivity typically cannot be ensured. In such networks, access to centralized registries (e.g., DNS or global object registries) is intermittent at best. When functionalities based on such registries are needed (e.g., entity resolution or entity linking), they have to be emulated or replaced within the local network, and then possibly integrated to the centralized infrastructure when the link to the wide-area network is reestablished.

We can for example imagine a XO laptop connected to a server (e.g., at school) downloading some Web pages or Wikipedia articles. Afterwards, the laptop is moved to a different location with no Internet connectivity but with the possibility to connect to other XO laptops. This scenario enables the sharing of previously downloaded documents with others who had no possibility to obtain them from the Internet, and the local publication of new entities.

In this context, it is often important to identify entities in all documents to enable entity-centric document aggregation, semantic of faceted search. Such aggregations may, for example, support learning applications that present the set of documents users should read when they want to learn about a specific entity (e.g., "Malaria").

Extraction entities from HTML text may be performed automatically by tools running on the XO laptops or even manually exploiting crowdsourcing, which can address the problem of limited computational resources available [4]. After an entity occurrence is identified in the text, it has to be uniquely identified by associating it with the right entity ID in order to foster automated processing. XO users can also create new entities themselves (e.g., through data acquisition, document authoring or document enrichment), which then should be propagated and shared to the rest of the local community. For those different tasks, an entity registry containing a relevant set of entity descriptions and identifiers is necessary to streamline and support all data-driven applications in ad-hoc networks.

2 Entity Registries Today

There already exist many solutions to resolve entity names and/or get structured information about entities. One example of entity registry has been proposed in the context of the Okkam project[7], where the envisioned system stores a number

[6] `http://linkeddata.org/`
[7] `http://www.okkam.org/`

of entity profiles which can be accessed via keyword or structured queries. More recently, the popularity of Linked Data made it possible to connect large entity datasets and to make them accessible via SPARQL endpoints. Additionally to these, we can imagine the adoption of well established platforms like, for instance, DNS or DOA and to extend such technologies to entity registries.

2.1 DNS

The Domain Name System (DNS) is the system used on the Internet to resolve domain names to their corresponding IP addresses. Domain resolution works in a hierarchical manner; the top of the domain name space is served by so-called *root* name servers, pointing to *authoritative* name servers maintaining authoritative information for the top-level domains (a.k.a. "TLDs", such as ".ch" or ".com"). The authoritative name servers responsible for the TLDs point in turn to further name servers, responsible for second-level domains (e.g., "unifr.ch"), and so on and so forth to process each domain name label iteratively until the last iteration, which return the IP address of the domain name queried. In practice, domain names are often cached at various levels, for instance at the client-side, or at the level of the DNS server provided by the Internet Service Provider in order to limit the load on authoritative DNS servers.

Though originally not designed for this purpose, it is be possible to extend the current DNS infrastructure to create a full-fledged entity registry. In that context, we recently suggested an extension of the DNS [3] to serve authoritative metadata about Internet domains, leveraging both the DNS Text Record field ("DNS TXT") and new cryptographic features ("DNSSEC").

2.2 DOA

The aim of the Digital Object Architecture (DOA)[8] is the management of digital entities over potentially very long timeframes. There are three distinct components in DOA:

- the Resolution System (Handle System)
- the Digital Object Repository (DORepository)
- the Digital Object Registry (DORegistry)

The principal function of the Handle system is to map known identifiers into handle records, containing useful information about the digital object being identified (e.g., IP address, public key, URL etc.). Every identifier has two parts: a naming authority (or prefix) and a unique local name under the naming authority suffix, separated by "/" (e.g. "10.1045/january99-bearman").

The collection of all local names defined under a certain prefix defines the local handle namespace under that prefix (something similar to a root zone in the case of DNS). All the local namespaces (i.e., all prefixes) define the handle namespace and a prefix can be considered as a top level domain. More namespaces for Local Handle Services (LHS) can be defined in a hierarchical fashion

[8] `http://www.cnri.reston.va.us/doa.html`

under the Global Handle Registry (GHR), thus the Handle system provides a hierarchical service model.

The DORegistry provides services like browsing, searching, repository and federation for collections of digital objects that can be distributed across multiple sites including other DO Registries. A DO registry may manage metadata of objects from a certain repository. Another possibility is managing both metadata and actual digital object content stored by the registry, and a third scenario is managing metadata from multiple repositories. The DO Registry can be set for different types of metadata schemata and can be customized to provide different search, federation, handle registration, event management and other services.

The most popular application of this system is the use of Digital Object Identifiers (DOIs) to identify digital versions of written publications (e.g., scientific articles). Such identifiers, by means of an ID resolution, will lead not only to the digital object but also to its metadata. The important benefit of using DOIs are persistent citations (i.e., the location of the digital object may change over time but the identifiers will remain the same and its resolution will lead to the new location).

2.3 Linked Data

The Linked Data movement has been pushing towards publishing and interlinking public data in standard formats, which enables the automated discovery, management and integration of structured resources online. The adopted technology is based on HTTP URIs and RDF. The resolution of an entity given its identifier boils down to three steps in that context:

1. discovering the IP address where the HTTP URI is supposed to be hosted (for example using the DNS)
2. contacting the corresponding server and negotiating the content (e.g., to serve a human-readable version of the RDF data if the client is a Web browser)
3. retrieve the structured description of the entity over HTTP.

This process is commonly called entity *dereferencing* since it is similar to general URI dereferencing on the Web[9].

2.4 The Entity Name System

In the context of the Okkam project, the Entity Name System (ENS) [2] has been proposed. It is defined as a service to resolve entity names to their global identifiers (called Okkam IDs). This is made available thanks to a repository of entity profiles described as a set of attribute-value pairs, and a mix of matching components that select the correct identifiers for an entity request which may be submitted in the form of a structured (i.e., attribute-value) or unstructured (i.e., keyword) query.

[9] `http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14`

3 Limitations of Current Entity Registries

The existing data and entity registries are a critical asset for many Internet applications. However, all current architectures present limitations when we consider a situation with limited network connectivity or ad-hoc environments. If we imagine networks of computers with no connectivity to external Internet resources, it becomes clear that DNS-based entity registries typically do not work properly as only partial information will be cached in local and accessible DNS nodes. Moreover, data in the cache may not be up-to-date as DNS nodes may not frequently communicate with the rest of the DNS infrastructure. Similarly, clients may not be able to resolve DOI if the servers of the naming authority which issued such IDs is not reachable through the network. Entity registries like the Okkam ENS or LOD end-points are all based on centralized solutions which limit their reliability: in ad-hoc environments such central resolution points may or may not be reachable at a given point in time.

For the above mentioned reasons, we claim that a decentralized solution enriched with full replication of some seed content would be a better approach for an entity registry in ad-hoc environments. A decentralized system based, for example, on structured P2P networks can provide better connectivity thanks to cheap P2P communication and can tolerate the situation in which registry servers are not constantly available. Thus, in the following section, we envision solutions that consist of multiple distributed registry instances to optimize the availability of entity-registries in the ad-hoc environments.

4 Approaches to Downscale Entity Registries

We sketch out below three possible avenues that we have considered for downscaling entity registries in ad-hoc environments. Our solutions are based on IPV6, structured P2P and full replication technologies.

4.1 IPV6 Addresses

For many, the first solution to our problem may be provided by IPV6 technologies. IPV6 guarantees directly-resolvable IP addresses, irrespective of the network topology (e.g., rending NAT irrelevant). Hence, local entities could be made available using local IP addresses only. One could even envision some address ranges reserved for such a use.

While practical and simple, we however think that solutions directly built on bare IPV6 technologies are limited for two reasons. First, we are missing the indirection layer offered by the DNS or similar registries, which is often essential when migrating or evolving entities (migrating a resource to another IPs would require setting up some redirection mechanism between the previous identifier and the new one). Second, from a networking perspective, this would boil down to mixing two separate layers, namely the networking and the application layers, which would certainly never be endorsed by the central networking companies and bodies.

4.2 Structured P2P Proxies

P2P technologies are another potential solution to our problem. Distributed Hash-Tables (DHTs), such as Chord[10] or our P-Grid system [1], provide decentralized, scalable hash-table-like functionalities that could be used to store entity identifiers as well as related meta information in ad-hoc environment. Through dynamic load-balancing and replication, those networks provide fault-tolerant and efficient networking primitives where arbitrary requests can typically be resolved in $O(log(N))$ messages, where N is the number of nodes in the P2P network, from any entry point to the network.

P2P technologies have been proposed in the past to enhance or supplant DNS infrastructures[11], most often to provide an alternative to ICANN or to support P2P file exchange. Such efforts had limited success so far. We think that the CoralCDN [5] system, in particular, is relevant to our scenario, since it takes advantage of highly efficient P2P mechanisms (P2P DNS and distributed sloppy hash tables) to create P2P content distribution networks. It however suffers from several severe limitations in our context, including some reliance on high-bandwidth and wide-area connectivity and the lack of any mechanism to serve structured entity content.

One could hence consider a DHT-based CDN as a starting point to solve our problem, and enhance the infrastructure with a native entity storage system (such as our recent dipLODocus system [7]), and with semi-structured capabilities (e.g., supporting declarative queries). Using such P2P infrastructures, we could for example explore the best possible way to support both entity publication and entity search in ad-hoc networks.

4.3 Entity Nucleus and Lazy Replication

Even though supporting a full-fledged entity registry in ad-hoc settings is a necessity, there are many cases where some of the nodes might connect to centralized infrastructures intermittently. Thus, we believe that it is also essential to be able to cache authoritative or centralized information, and to be able to dynamically synchronize data with such infrastructures.

Depending on the context of the application, some core nucleus of the entity data can be identified (e.g., DBPedia entity data for LOD). In such a case and if the entity nucleus can be pre-installed on each node, then many of the entity operations can be resolved locally without resorting to any third-party infrastructure.

Networked search and updates, however, still require distributed mechanisms to be resolved. If such operations are deemed relatively infrequent, then semi-structured P2P technologies like those described above can be applied: both distributed queries and updates can for instance be (lazily) propagated or broadcasted across the network at regular intervals.

---

[10] `http://pdos.csail.mit.edu/chord/`

[11] see for instance `http://blogs.computerworld.com/17444/p2p_dns_to_take_on_icann_after_us_domain_seizures`

5 Conclusions

Current entity registry solutions are often based on global hierarchies or centralized online directories. Such solutions are inapplicable to many contexts, including ad-hoc networks and environments that have limited access to a wide-area connection. In this paper, we described some of the key problems related to using current entity registries in an ad-hoc context and suggested a few possible alternatives. Three potential solutions were specifically sketched, based on IPV6 technologies, scalable and structured P2P technologies, and (lazy) content replication. We now plan to implement, test and combine both current entity registry solutions and our new alternatives to determine in practice which architecture is most useful given a specific ad-hoc environment. As a start, we plan to focus on the OLPC XO context to provide a working solution and enable Open Data and Linked Entity applications for the billions of people who are currently cut-out of wide area networks.

6 Acknowledgment

This work was supported (in part) by the Swiss National Science Foundation under grant number PP00P2_128459.

References

1. Karl Aberer, Philippe Cudré-Mauroux, Anwitaman Datta, Zoran Despotovic, Manfred Hauswirth, Magdalena Punceva, and Roman Schmidt. P-grid: A self-organizing structured p2p system. *ACM SIGMOD Record*, 32(3), 2003.
2. Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and Antonio Mana. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, number CSS-ICSC 2008-4-28-25 in CSS-ICSC, pages 554–561. IEEE Computer Society, August 2008.
3. Philippe Cudré-Mauroux, Gianluca Demartini, Djellel Eddine Difallah, Ahmed Elsayed Mostafa, Vincenzo Russo, and Matthew Thomas. A Demonstration of DNS^3: a Semantic-Aware DNS Service. In *ISWC 2011*.
4. Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudre-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In *International World Wide Web Conference (WWW)*, 2012.
5. Michael J. Freedman, Eric Freudenthal, and David Mazières. Democratizing content publication with coral. In *Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation - Volume 1*, NSDI'04, pages 18–18, Berkeley, CA, USA, 2004. USENIX Association.
6. Christophe Guéret, Stefan Schlobach, Victor De Boer, Anna Bon, and Hans Akkermans. Is data sharing the privilege of a few? Bringing Linked Data to those without the Web. In *ISWC 2011 - Outrageous Ideas*.
7. Marcin Wylot, Jigé Pont, Mariusz Wisniewski, and Philippe Cudré-Mauroux. dipLODocus[RDF] - Short and Long-Tail RDF Analytics for Massive Webs of Data. In *International Semantic Web Conference*, pages 778–793, 2011.