Effective digital object access and sharing over a networked environment using DOIP and NDN

Cas Fahrenfort
System and Networking Lab
University of Amsterdam
the Netherlands
casfahrenfort@gmail.com

Zhiming Zhao
System and networking Lab
University of Amsterdam
the Netherlands
z.zhao@uva.nl

***Abstract*—FAIRness (findability, accessibility, interoperability and re-usability) is crucial for enabling open science and innovation based on digital objects from large communities of providers and users. However, the gaps among version control, identification and distributed access systems often make the scalability of data centric applications difficult across large user communities and highly distributed infrastructures. This poster proposes a solution for accessing and sharing digital objects over a networked environment using Digital object interface protocol (DOIP) and Named Data Networking (NDN).**

***Index Terms*—Digital Object Interface Protocol, Named Data Networking, Persistent Identifier, FAIRness**

I. Introduction

Nowadays, data are playing an increasingly important role in scientific research (e.g., for modelling complex environmental systems or for early warning upcoming disasters), and industrial innovations (e.g., for business service recommendation and multimedia distribution [1]). Being often collected by distributed sources (e.g. environmental observations from different geo-regions), data are evolving over time (e.g. collaborative contributions on media content driven by common community interests or data quality controls over continuous environmental monitoring) and are managed by different parties.

To develop a data application using data from distributed sources, one has to first effectively discover and find suitable data objects for the specific application purpose, and then access and integrate the discovered data objects in the customized processing pipeline. However, those data objects often do not have same level of details in the meta information, and lack globally resolvable names (or identifiers), which make the discovery, access and citation of those data objects difficult. Moreover, the evolution of the digital objects results in different versions, which often leads to complicated dependencies when those objects are used in creating new objects. Without rich contextual information of different versions, the (re)usage of data objects in a data workflow is often error prone. Finally, many data applications are often based on community collaboration, for example composing or collaborative editing of online education or training material using community contributed media. When the size of digital objects are very large, e.g., editing video content, the performance of the object sharing is critical for high quality of user experiences. Centralized storage of data objects often creates performance bottlenecks for distributed users.

The work is partially supported by EU H2020 ENVRIPlus and ENVRI-FAIR projects.

Advanced data infrastructures are thus required to effectively manage the dynamic evolution (versioning), identification and citation (globally resolvable names) of digital objects, and enable their high find-ability, accessibility, interoperability and re-usability (namely FAIR) among distributed communities. However, lots of existing data management systems have originated from early legacy systems, e.g., environmental observation stations, and lack a global data identifier centered design for data services. This poster discusses how community standards like the Digital Object Interface Protocol (DOIP) [2] and advanced Cloud and networking technologies like Information Centric Networking and, more specifically, Named Data Networking (NDN) [3] can enhance the FAIRness of a data infrastructure.

II. Challenges and related work

To seamlessly integrate the management services for versioning control, publishing, discovery and distribution of digital objects, globally resolvable and persistent identifiers of the digital objects are crucial. Moreover, the data infrastructure must 1) facilitate data discovery and identification, 2) guarantee available metadata for all digital objects and 3) support resolution of persistently identified data. Furthermore, the distribution system should 1) have reasonable performance, 2) be scalable with the amount of traffic occurring on the network and 3) be easily integratable for both organizations and consumers with legacy systems.

A. State of the art

The Digital Object Interface Protocol (DOIP) defines 1) an Identifier/Resolution system for allotment and resolution of Persistent Identifiers (PIDs), 2) a Repository System for storing digital objects and 3) a Registry System, which is a search-able metadata resource for all digital objects stored in the Repository System.

Named Data Networking (NDN) is an Information Centric Networking (ICN) solution for routing objects through a net-

work based on object name instead of location. Additionally, NDN caches passing objects at each router hop to increase network efficiency.

NDN-as-a-service for PID (NaaS4PID) [4] is an extension of NDN which allows it to interoperate with PIDs. NaaS4PID adds an additional system layer which translates PIDs to NDN names and optimizes and manages the virtual NDN overlay in Cloud or other e-infrastructures.

B. Gap analysis

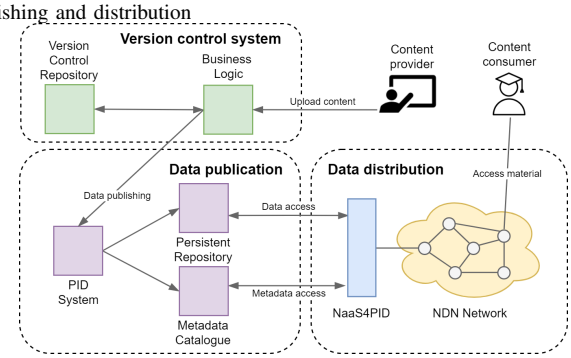
The technologies mentioned in Section II-A come from different communities, and can not be seamlessly integrated. NDN and DOIP data storage have different naming schemes, and require an additional service (NaaS4PID) to interoperate. Current version control applications, such as Git [1], store data in their own specialized formats and repository structures, and the same applies to many existing legacy systems.

III. Architecture and Technical Considerations

A. A DOIP centered versioning, publishing and distributing data management solution

Figure 1 shows a preliminary architecture for persistently publishing digital objects from an organization's legacy system. Following from the workflow in Figure 2, it is divided into three distinct parts: version control system, data publication and data distribution. Content created by community content providers is processed and stored (in its internal repository) by the version control system in any form that is required by its specific use case. Once submitted content has been approved, it is published to persistent data and metadata repositories by any PID system. From there, the data can be discovered and accessed by content consumers through the NDN network and NaaS4PID service.

Fig. 1. A DOIP centered architecture for digital object version control, publishing and distribution



B. Technical considerations

This architecture is generic in the sense that any existing legacy system can adopt it. The system only needs to be expanded with functionality to publish the data to the PID system, including a mapping between existing data structure and a (meta)data structure fitting to the PID system.

[1] www.git-scm.com

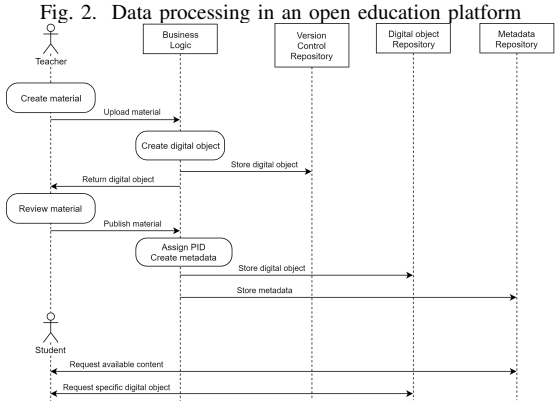
In the architecture as shown in Figure 1, only a single data provider is present on the network, since NaaS4PID can currently only support one provider. Additional content providers could offer large community benefits, such as third parties independently creating education material. However, there are as of yet no solutions that can provide such functionality.

Furthermore, it is assumed that data publication and PID assignment/resolution is handled by the platform itself. However, having separated resolution/publication from the internal business logic could allow for benefits, such as allowing other organizations to publish data to the same repositories.

IV. Case Studies

By integrating with a system such as described in Section III, a community platform for providing educational content in an open manner, i.e. accessible and reusable by everyone, as well as adaptable to individual needs could enhance their data FAIRness. Figure 2 shows an example workflow for a teacher in an open education platform. Material is first created using internal systems logic and saved in an internal repository containing versioned data. A student can later discover and request content through publicly accessible (meta)data repositories.

Fig. 2. Data processing in an open education platform



V. Summary

This poster presents the ongoing work in enhancing FAIRness for legacy data management solutions. A DOIP centered digital object versioning, publishing and distribution solution is presented.

References

[1] Y. Mo, J. Chen, X. Xie, C. Luo and L. T. Yang, "Cloud-Based Mobile Multimedia Recommendation System With User Behavior Information," in IEEE Systems Journal, vol. 8, no. 1, pp. 184-193, March 2014.

[2] DONA Foundation, "Digital Object Interface Protocol Specification," DONA Foundation, 2018.

[3] L. Zhang and D. Estrin and J. Burke and V. Jacobson and J. D. Thornton and D. K. Smetters, "Named data networking (ndn) project," in Relatório Técnico NDN-0001, Xerox Palo Alto Research Center-PARC, vol. 157, pp. 158, Citeseer, 2010.

[4] S. Koulouzis and R. Mousa and A. Karakannas and C. de Laat and Z. Zhiming, "Information centric networking for sharing and accessing digital objects with persistent identifiers on data infrastructures," in 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 661–668, IEEE, 2018.