From Me to You: Peer-to-Peer Collaboration with Linked Data

Jan Kaßel and Dr. Thomas Koentges

Leipzig University, Germany

Abstract

In recent years, Digital Humanities' collaborative nature caused a wake digitally-native research practice, where interdisciplinary workflows commonly feed into centralized data repositories. Connecting these repositories, the W3C's Web Annotation specification builds upon Linked Data principles for targeting any Web resource or Linked Data entity with syntactic and semantic annotation. However, today's platform-centric infrastructure diminishes the distinction between institutional and individuals' data. This poses issues of digital ownership, interoperability, and privacy of data stored on centralized services. With Hyperwell, we aim to address these issues by introducing a novel architecture that offers real-time, distributed synchronization of Web Annotations, leveraging contemporary peer-to-peer technology. Extending the peer-to-peer network, institutions provide Hyperwell gateways that bridge peers' annotations into the common web. These gateways affirm a researcher's affiliation while acting as a mere mirror of researchers' data and maintaining digital ownership.

Keywords

linked data; peer-to-peer systems; annotation; real-time collaboration

I INTRODUCTION

Research projects in the Digital Humanities continuously establish new standards in digital workflows with novel tools. This development is facilitated by an increasing number of open source projects, available for free use and maintained by a large community of voluntary contributors. Commonly respected principles for managing data in research projects, such as the FAIR[1] principles (Findable, Accessible, Interoperable, Reusable), establish a theoretical framework for these workflows, complementing technical architecture. Furthermore, by incorporating end-to-end, persistent references with technologies such as Digital Object Identifiers (DOIs), contemporary digital publishing takes on these frameworks. Thereby, references to particular versions of data can be resolved via open archives such as Zenodo[2] and HAL[3].

Reaching this level of quality in digital workflows is a major achievement for academia. Yet, the way we currently treat our data repositories is, in general, negligent. Often archived on platforms running on remote hardware in far-away data centers, people are rarely able to *exactly* prove (that is, mathematically) *when* and *what* they contributed to a data repository. After all, at this point, digital data ownership is mostly given when data resides on one's personal computing device. Local availability of data is itself advantageous: In 2019, Kleppmann et al. coined the term *local-first applications* in reference to software that operates on locally stored data, as opposed to the popular approach of *thin clients*, where data is mainly pushed into servers running remotely. However, local-first applications further question the necessity of centralized services for real-time collaboration—take Google Docs—by asserting that even as data leaves one's devices, digital ownership can be proven by cryptographic means. This concept provides

---

[1] https://www.go-fair.org/fair-principles/
[2] https://zenodo.org/
[3] https://hal.archives-ouvertes.fr/

one major takeaway: Personal data doesn't necessarily need to be shared with third parties when collaborating with peers.

Applying this concept to research, we aim to pose a distinction between personal and institutional data in the context of annotation: Artifacts, such as classic texts or digitized manuscripts, belonging into the institutional domain, are being stored remotely. Annotations on these sources, however, can be stored on an individual's device. Our work describes such workflows and, by introducing a system called Hyperwell, provides an architecture that leverages contemporary peer-to-peer technology to realize local-first, distributed, and collaborative annotation environments.

II COLLABORATION IN THE DIGITAL HUMANITIES

In the wake of *Web 2.0*—a technological development that emerged after the dot-com bubble burst in 2001—the internet became more open and accessible to the general public. Collaboration in research, and particularly in the Digital Humanities, adopted this trend, as the web tended towards the social sharing of resources on websites (Davidson, 2008). Two advancements then materialized within the Digital Humanities: as human expression became increasingly digital, Humanities' workflows turned digital, too; subsequently, information on the past has been digitized, emphasizing digital practice within the Humanities.

2.1 Eposing History

The FAIR principles consist of four imperatives for digital workflows—Findability, Accessibility, Interoperability, and Reusability—, and have been rapidly adopted among scholars. Some best practices continue to go even further by explicitly recommending version control and the recording of additional metadata (Nowogrodzki, 2020). Maintaining a log of changes is a convenient, self-documenting step in end-to-end digital workflows.

Over the years, the Git[4] version control system has been established as a popular tool in software development, as it allows users to write code in a change-based workflow: Collections of changes are bundled into a commit, which again is written into an immutable log and then distributed to others. This log, also called an append-only log, makes it straightforward to maintain a persistent history of changes and enables time-traveling among past states (Kreps, 2013).

2.2 Web Annotation

With pen-and-paper annotation, one's expression is basically unlimited: lines, highlightings, and words can connect any visible artifact. Transforming such free interaction into digital processes can entail an unforeseeable complexity. However, as Marshall noted while analyzing a collection of used students' textbooks, a majority of their annotations fit into a categorization of six distinct schemes, varying in their focus and level of detail from simple marginal symbols up to complex sidenotes (Marshall, 1997).

Published by the World Wide Web Consortium (W3C), the Web Annotation specification is built upon an ontology that is expected to be as versatile as possible, covering annotation schemes such as the above by providing varieties of target selection mechanisms and content types (Sanderson et al., 2013). The fundamental parts of an annotation within this ontology are its target (a web resource) and its body (an entity). By using semantic expressions, an annotation body can convey complex structures, such as their own motivation (for example, editing a passage) and their relations to other objects (for example, referring to a particular person). Such
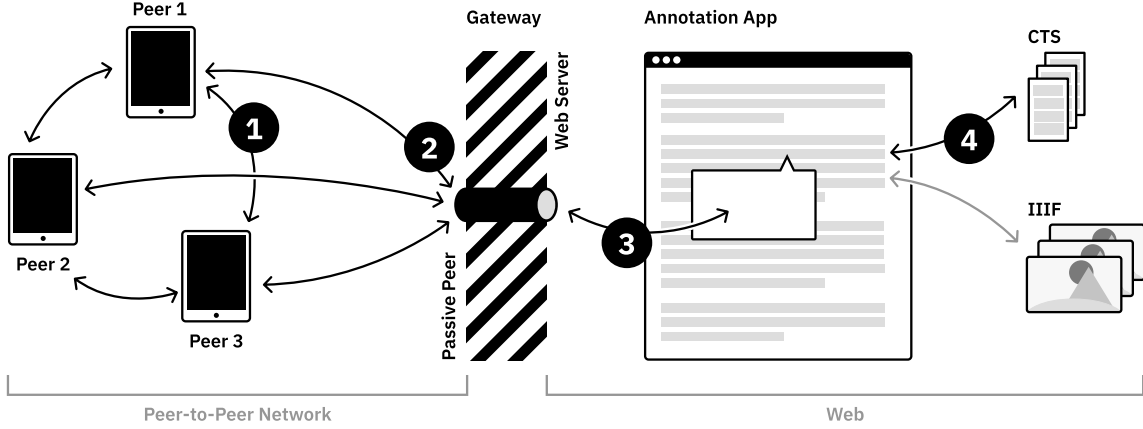
---

[4] https://git-scm.com/

Figure 1: The Hyperwell architecture. Peers exchange their notebooks in real-time (1). Gateways, run by institutions, archive selected notebooks and bridge them into the web (2). Web applications can access annotations via gateways, as they implement the Web Annotation protocol (3). These applications can load canonical resources via services such as CTS or IIIF (4).

structures could meet the semantic requirements of digital datasets in the Humanities, such as gazetteers in the Classics.

III   PERSONAL ANNOTATIONS WITH HYPERWELL

With our work on personal annotation, Hyperwell, we aim to address several concerns at once. Hyperwell imposes a clear architectural distinction between *personal* and *institutional* work enabled by leveraging peer-to-peer technology. We introduce the notion of a digital, personal notebook, residing on individuals' devices. Such a notebook is a collection of annotations on a single resource and complies with the Web Annotation data model[5].

Figure 1 outlines the architecture of Hyperwell. Each notebook can be shared directly with selected peers. By exposing an immutable, append-only log of all changes made to it, each version of a notebook can be referenced individually and facilitates archiving. Special requirements on the data types used allow the realization of real-time collaboration without the need for a central authority to solve merge conflicts between multiple peers. Finally, end-to-end referencing of such workflows becomes possible, as each version of a notebook can target canonical resources, such as passages via Canonical Text Services (CTS) or multimedia resources via IIIF[6].

3.1   Bridging the Gap with Gateways

Classic notions of *pure* P2P networks impose the limitation that within such networks, resources are shared by nodes homogeneously, as Schollmeier (2002) notes. This homogeneity ensures that data within these networks is genuinely decentralized. Network architects occasionally break with this homogeneity for various reasons, for instance, to improve performance or to ensure interoperability with other networks.

In a patent from 2010, Matsubara and Miki sketch how such interoperability could be realized by describing a gateway between a peer-to-peer network and the web. Similar to the gateway pictured in Figure 1, it acts as an intermediary between individual peers and requests from web clients. We implemented a gateway for Hyperwell that acts both as a passive peer within the

---

[5]https://www.w3.org/TR/annotation-model/
[6]https://iiif.io/

3

peer-to-peer network and as a server for web clients (Kaßel, 2020). The gateway implements the Web Annotation protocol[7] for compliance with environments using the Web Annotation specification.

3.2 Institutional Archival of Notebooks

Along with the aforementioned advantageous prospects of peer-to-peer networks, there is also a particular risk: If the device serving a certain resource is disconnected from the network, this resource is not available to other peers. In the academic context, however, and especially for published datasets or collaborative work, continuous data availability is crucial.

Supporting nodes can improve data availability and, hence, the overall quality of a peer-to-peer network. By *pinning* or *seeding* resources, such nodes can mirror other peers' data without actually causing any increase in the centralization of resources—quite the contrary, this affects a further distribution, while as volatile infrastructure, supporting nodes maintain original ownership (Benet, 2014; Legout et al., 2007). Within the architecture of Hyperwell, such a supporting infrastructure operates in the interest of institutions: They can not only ensure data availability for affiliated researchers, but also archive repositories in their entirety, as each repository contains its complete history.

3.3 Real-Time Collaboration

Merging changes from different sources is rarely a trivial task, especially within a distributed environment. Collaborative applications like Google Docs and Trello use a central authority—their servers—to solve these conflicts, but this requires a continuous internet connection from clients and relies on this authority. However, contemporary technology can already address this challenge: Conflict-Free Replicated Data Types (CRDTs) impose several theoretical merging strategies on data in order to obtain consistent and fail-safe merging in distributed networks (Kleppmann et al., 2019).

To realize this functionality in Hyperwell, we utilize *hypermerge*[8], a library that combines several distributed technologies for realizing real-time, peer-to-peer collaboration:

- *automerge*[9], a JavaScript-based CRDT implementation,
- *hypercore*[10], a distributed append-only log, allowing us to immutably store and distribute changes on a dataset, and
- *hyperswarm*[11], a stack of distributed networking technologies for establishing connections between peers without centralized discovery.

Both hypercore and hyperswarm are being actively developed by the Dat Protocol project, a non-profit group building a distributed data-sharing protocol (Robinson et al., 2018).

IV OUTLOOK

With Hyperwell, we presented a collaborative, peer-to-peer annotation system that builds upon contemporary peer-to-peer technology and implements the Web Annotation specification. Hyperwell aims to address issues concerning data privacy and digital sovereignty in Digital Humanities research, facilitated by centralized platforms. Establishing a separation of personal and

---

[7]https://www.w3.org/TR/annotation-protocol/
[8]https://github.com/automerge/hypermerge
[9]https://github.com/automerge/automerge
[10]https://github.com/mafintosh/hypercore
[11]https://github.com/hyperswarm/hyperswarm

institutional data, we present a gateway service for Hyperwell, bridging personal, distributed annotation notebooks into the web for interoperability with existing annotation environments.

Further work will focus on evaluating the presented architecture and its entailing collaborative workflows. To provide the necessary prototype application for such an environment, we aim to build two applications: First, a local-first notebook application that will manage annotations by storing them locally and indexing them for local search. Second—for conducting actual usability and workflow evaluation—a proof-of-concept, collaborative environment that integrates canonical resources such as CTS passages or multimedia IIIF manifests.

Bibliography

Benet, J. (2014). IPFS - content addressed, versioned, p2p file system. *arXiv:1407.3561 [cs]*, arxiv 1407.3561. Retrieved May 16, 2019, from http://arxiv.org/abs/1407.3561

Davidson, C. N. (2008). Humanities 2.0: Promise, perils, predictions. *PMLA*, *123*(3), 707–717. https://doi.org/10.1632/pmla.2008.123.3.707

Kaßel, J. (2020, January). *Hyperwell gateway* (Version v0.2.0). Zenodo. https://doi.org/10.5281/zenodo.3631524

Kleppmann, M., Wiggins, A., van Hardenberg, P., & McGranaghan, M. (2019). Local-first software: You own your data, in spite of the cloud, In *Proceedings of the 2019 ACM SIGPLAN international symposium on new ideas, new paradigms, and reflections on programming and software - onward! 2019*. the 2019 ACM SIGPLAN International Symposium, Athens, Greece, ACM Press. https://doi.org/10.1145/3359591.3359737

Kreps, J. (2013, December 16). *The log: What every software engineer should know about real-time data's unifying abstraction* [LinkedIn engineering]. Retrieved January 30, 2020, from https://engineering.linkedin.com/distributed-systems/log-what-every-software-engineer-should-know-about-real-time-datas-unifying

Legout, A., Liogkas, N., Kohler, E., & Zhang, L. (2007). Clustering and sharing incentives in BitTorrent systems. *ACM SIGMETRICS Performance Evaluation Review*, *35*(1), 301. https://doi.org/10.1145/1269899.1254919

Marshall, C. C. (1997). Annotation: From paper books to the digital library, In *Proceedings of the second ACM international conference on digital libraries - DL '97*. the second ACM international conference, Philadelphia, Pennsylvania, United States, ACM Press. https://doi.org/10.1145/263690.263806

Matsubara, D., & Miki, K. (2010, August 3). *Method and apparatus for peer-to peer access* (U.S. pat. 7769881B2). Retrieved January 26, 2020, from https://patents.google.com/patent/US7769881B2/en

Nowogrodzki, A. (2020). Eleven tips for working with large data sets. *Nature*, *577*(7790), 439–440. https://doi.org/10.1038/d41586-020-00062-z

Robinson, D. C., Hand, J. A., Madsen, M. B., & McKelvey, K. R. (2018). The dat project, an open and decentralized research data tool. *Scientific Data*, *5*, 180221. https://doi.org/10.1038/sdata.2018.221

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013). Designing the w3c open annotation data model, In *Proceedings of the 5th annual ACM web science conference on - WebSci '13*. the 5th Annual ACM Web Science Conference, Paris, France, ACM Press. https://doi.org/10.1145/2464464.2464474

Schollmeier, R. (2002). A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications, In *Proceedings first international conference on peer-to-peer computing*. First International Conference on Peer-to-Peer Computing, Linkoping, Sweden, IEEE Comput. Soc. https://doi.org/10.1109/P2P.2001.990434