Hyperwell: Local-First, Collaborative Notebooks for Digital Annotation

Jan Kaßel

A thesis presented for the degree of
Master of Science

Supervised by:
Dr. Thomas Köntges
Prof. Gregory Crane

Leipzig University, Germany
January 2015

I, Jan Kaßel confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam et turpis gravida, lacinia ante sit amet, sollicitudin erat. Aliquam efficitur vehicula leo sed condimentum. Phasellus lobortis eros vitae rutrum egestas. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec at urna imperdiet, vulputate orci eu, sollicitudin leo. Donec nec dui sagittis, malesuada erat eget, vulputate tellus. Nam ullamcorper efficitur iaculis. Mauris eu vehicula nibh. In lectus turpis, tempor at felis a, egestas fermentum massa.

Acknowledgements

Interdum et malesuada fames ac ante ipsum primis in faucibus. Aliquam congue fermentum ante, semper porta nisl consectetur ut. Duis ornare sit amet dui ac faucibus. Phasellus ullamcorper leo vitae arcu ultricies cursus. Duis tristique lacus eget metus bibendum, at dapibus ante malesuada. In dictum nulla nec porta varius. Fusce et elit eget sapien fringilla maximus in sit amet dui.

Mauris eget blandit nisi, faucibus imperdiet odio. Suspendisse blandit dolor sed tellus venenatis, venenatis fringilla turpis pretium. Donec pharetra arcu vitae euismod tincidunt. Morbi ut turpis volutpat, ultrices felis non, finibus justo. Proin convallis accumsan sem ac vulputate. Sed rhoncus ipsum eu urna placerat, sed rhoncus erat facilisis. Praesent vitae vestibulum dui. Proin interdum tellus ac velit varius, sed finibus turpis placerat.

Contents

List of figures

List of tables

Abbreviations

- DHT
- CRDT
- P2P
- HTML

API	**A**pplication **P**rogramming **I**nterface
JSON	**J**ava**S**cript **O**bject **N**otation

Chapter 1

Introduction

1.1 Collaboration in the Digital Humanities

The Humanities are one of the world's oldest research areas, gathering and discovering knowledge about our own history, being, and society (Davidson, 2008). (Transition to digital humanities and, hence, a transition from analog to digital artifacts).

With the wake of technological development that happened after the bust of the dot-com bubble in 2001, the internet became more open and accessible to the common people, as opposed to purely experts (Davidson, 2008). Collaboration, again, has been facilitated by this change, as the web tended towards social media, sharing of resources, and semantic description of data.

Not only has the subject's *habitus* adopted this transformation: Most of humanities' contemporary expression is of digital nature—take modern literature, for instance, where Silvia Hartmann populates the progress on her work in real-time via Google Docs (Kirschenbaum, 2016).

But even more, the scholarly methodologies considering the matters of the pre-digital ages changed. With the rise of Natural Language Processing (NLP) of commonly approachable Machine Learning, (more about what's currently possible, and since when. Terras et al. (2016) might be useful!).

Physical artifacts became digital resources, analogue workflows became digital ones. The digital nature of the contemporary humanist's work suddenly demanded additional, extensive knowledge of gathering, analyzing, and maintaining data repositories. Best practices were formed, namely the FAIR principles (Wilkinson et al., 2016): Findable, Accessible, Interoperable, and Reusable data.

Being an umbrella for various disciplines, the Humanities historically embraced academic collaboration to a great extend (Siemens, 2015).

Something on Linked Data (Simon et al., 2015, 2017)?

Something on annotation (Kahan et al., 2002; Marshall, 1997, 1998; Sanderson et al., 2013)?

In the wake of the new web, the wake of technological development that happened after the bust of the dot-com bubble in 2001, the internet became more open and accessible to the common people, as opposed to purely experts (Davidson, 2008). Collaboration, again, has been facilitated by this change, as the web tended towards social media, sharing of resources, and semantic description of data.

1.2 Recogito

Using annotation to enhance source material with additional information is a practice commonly used in the Humanities, as it provides an environment for connecting sources with one's personal thoughts. The emergence of digital tools in the Humanities then brought not only the likes of computational methods, but also made the social aspects of the internet more approachable to non-experts (Davidson, 2008).

This caused two particular implications. Collaboration, an artifact of particular importance in the interdisciplinary nature of Humanities research, enabled working on the same resources with multiple people involved, exchanging their findings—though, digitally—in real-time. Furthermore, the rise of the Semantic Web with Linked Data actually imposed semantic meaning on data itself: Digital entities could reference each other and describe their relations in more depth.

Recogito, a project by the Pelagios Commons, takes on to leverage Linked Data for spatial annotation. Digital gazetteers, such as the Digital Atlas of the Roman Empire (DARE, https://dh.gu.se/dare/), can be imported into Recogito and tagged from within a resource like, for example, the Iliad, exposing semantic relationships (Simon et al., 2017). A separate map view on Recogito plots these relations on a map and allows for exploring them visually. Annotations on such a resource can then be shared with fellow users, providing opportunities for further contributions.

Exploring historic artifacts in such ways provides new ways of perceiving historic information. Besides being beneficial, this could even more benefit educational programs in the Classics. We set out to explore these prospects in a user testing session, involving students at Furman University in a semi-controlled environment …

The takeaways of this publication are two-fold: First, we present two modular dataset created by the participants during both sessions on the mentioned resources—the Iliad's Catalog of Ships, and the Tabula Tabula Peutingeriana. They contain georeferences for many places as well as annotations relating to people, events, and general remarks. These datasets are available in various formats and can easily be imported into Recogito.

Second, we present an overview of the survey and analytics results received during the sessions, focusing on the perceived benefits of spatial annotation and collaboration, as well as the general user experience.

1.3 Digital Authoring with Hypertext

The way humanity treats its expression sets implications on how research can reflect on these. Marshall (1997) introduced a fundamental notion on this in regard to books and readers' markings: The physical representation of a book bears the respective physical limitations of adding further information to paper-based text—being it highlights or marginal notes, for instance. (…)

The digital medium, however, lifts those physical limitations. Almost encouragingly, adding annotations to a digital text is just a matter of switching bits from zero to one. (…)

1.4 Peer-to-Peer Systems

1.5 Outline of This Thesis

This is a brief outline of what went into each chapter. **Chapter 1** gives a background on duis tempus justo quis arcu consectetur sollicitudin. **Chapter 2** discusses morbi sollicitudin gravida tellus in maximus. **Chapter 3** discusses vestibulum eleifend turpis id turpis sollicitudin aliquet. **Chapter 4** shows how phasellus gravida non ex id aliquet. Proin faucibus nibh sit amet augue blandit varius.

Chapter 2

Related Work

2.1 Digital Humanities Infrastructure

How are resources treated in the Digital Humanities? Canonical text systems such as CTS have been available for a couple of years, and IIIF is currently emerging and becoming more popular among GLAM institutions. Give an overview of some Digital Humanities tools, such as Recogito or Ugarit, to emphasize the distinction between institutional and personal research data.

2.2 Hypertext and Hypermedia Systems

2.3 Web Annotation and Linked Data

2.4 Peer-to-Peer Technologies

Describe the fundamental technologies first: Append-only logs, Distributed Hash Tables (DHTs), Conflict-Free Replicated Data Types (CRDTs). Introduce contemporary systems that leverage these fundamentals: IPFS and Dat, and maybe some previous attempts such as Gnutella and Skype. Blockchain could provide a good take on emerging high-tech, and federated systems also are of interest, if that doesn't extend the section's boundaries too much. # Hypertext Annotations

2.5 Lorem Ipsum

The way we treat annotations today is wrong, and by considering how tools for annotating PDFs and websites work we can learn about the issues. Coming from the Related Work section, this chapter should give a theoretical introduction into issues of text theory and how annotations fits in there. By leveraging a comparison of OHCO and Hypertext (or, Ted Nelson's Xanadu), we should derive an architecture for annotations, and be able to show how annotations work in the Web's notion of hypertext.

Chapter 3

User Testing

3.1 Lorem Ipsum

Lorem ipsum.

Chapter 4

Gateways in Distributed Systems

> This will most likely be an argumentation why we need P2P gateways when working with P2P data in academia: Many platforms and tools are built with web technologies and hence are subject to the quasi-centralized architecture of the HTTP web.

Distributed P2P systems function fundamentally different from the classic client/server architectures (distributed governance figure?). The fundamental difference is explained by the treatment of data: In architectures following the established client-server distinction, such as HTTP, servers hold a monopoly of the contained data while clients request parts of this data on demand. This provides several benefits for businesses: They are able to govern the singular source of their services' data by properly "owning" it. This means, businesses are effectively controlling aspects such as data availability, access to data, its versioning, and basically any kind of operation on it, ensuring commercial exploitation. (Something on providing guaranteed uptime, data backups, etc.).

In P2P systems, this power over data is distributed. The distinction of clients and servers is being blurred as the centralization of governance is diminished: Clients become servers, forming a collection of alike peers, that provide and at the same time request data. Considering "the data" a system operates on as a database (with support for querying and mutation), in these kind of distributed systems, this database is distributed, sometimes even fragmented.

This poses many questions when conceiving P2P architectures: Which parts do work well centralized? Which functionality does effectively when being distributed? How can certain control structures be realized?

One trade-off of theoretically "pure" P2P systems is, considering all data is exchanged between genuine peers, that each peer is running on commodity hardware—regular consumer devices. Especially in these days, where an increasing number of our interactions with the digital world occurs via handheld devices such as smartphones, their lack of processing power compared to

the enormous computational resources of a dedicated cluster is troublesome. Yet, with the wake of the more mature, "smarter" P2P systems, these inequalities were to be addressed. Skype, for instance, as research by Guha (2015) showed, analyzed peers' network performance and promoted particular peers to supernodes. These supernodes "maintain an overlay network network among themselves" (Guha, 2015, p. 2) and effectively outbalance the weaknesses of less powerful peers (Chawathe et al., 2003).

(Textile cafés?).

In the following, I will describe two attempts at an implementation for a system that bears a critical burden: Realizing a distributed system that bridges its data into the web via HTTP. The question of where to put that bridge shapes the distinction between both attempts: With the first attempt described in section X, the "Thick" Peer, that bridging is provided from within each peer, effectively ensuring the realization of distributed, independent publishing of one's annotations. As I will lay out in the following, putting that much liability, and hence, network load, onto an independent peer, will quickly exhaust the given resources and hinder the scalability of this approach. With the second, more successful attempt presented in section Y, this liability is moved into institutional governance: While peers exchange their data within the P2P network, the task of bridging that data into the web is done by institutions who run quasi-centralized gateways. As tests showed, this attempt scales well with real-time updates, while individual peers are excused from responding to a growing number of HTTP requests.

4.1 Bridging Into The Web

Web applications leverage technologies planned, audited, and released by the World Wide Web Consortium (W3C). These technologies are known as *web technologies* and are commonly supported by web browsers such as Mozilla Firefox, Google Chrome, and macOS Safari. Web applications are a popular way of providing tools and services, as opposed to native applications executed directly by the user's operating system, due to three factors: * User Experience (UX): Websites are accessible by entering Uniform Resource Locators (URLs) such as `https://www.eff.org/`. Users don't have to manually download an application bundle and run it on their machine, as browsers download the application code and assets in-promptu. * Developer Experience (DX): Developers can choose from a variety of standardized, open technologies for realizing their applications: Building web documents with HTML, realizing complex business logics with JavaScript, * Business Benefits: …

Chapter 5

Gateways and P2P Systems

Each aspect of a P2P system bears implications for usability, data availability, and user emancipation: As described in the previous chapters, P2P networks can effectively use certain network structures to enforce power structures and hierarchies among peers.

Chapter 6

Design & Implementation of Hyperwell

6.1 Resource Exhaustion: A “Think” Peer

Describe issues with the first iteration of Hyperwell, where the Gateway API was residing in each and every peer.

```
mood = 'happy'
if mood == 'happy':
    print("I am a happy robot")
```

Alternatively, you can also use LaTeX to create a code block as shown in the Java example below:

Listing 6.1: Main.java

```
1 /**
2  *  Hello, world - example in Java.
3  */
4 public class Main{
5     // says hello to the world
6     public static void main(String[] args) {
7         System.out.println("Hello, world!");
8     }
9 }
```

6.1.1 Resource Discovery

6.1.2 Client SDK

6.2 Institutional Governance with "Hyperwell"

6.2.1 Gateway: Implementation of a Service for Archival and Institutional Exposition

6.2.2 Notebook: Implementation of a Local-First Annotation Application

If there's enough time to realize the local notebook application, write a small chapter about it here.

6.2.3 Adoption in Existing Environments

Explicate how our approach on adding Hyperwell support to the Recogito semantic annotation platform went.

6.2.3.1 A Standard Annotation Server

6.2.3.2 Adding Real-Time Collaboration Support

6.3 Conclusion

With many platforms involved—researchers, institutions, platforms, non-academic users—it's difficult to find the perfect solution suiting all their needs. The "Thick Client" approach presented first ensures an annotator's independence when publishing, but bears the quick exhaustion of their computational resources. The second approach, Hyperwell, performed well in testing due to a clear distinction of personal (individual) and institutional (centralized) computational resources, but takes the way of introducing quasi-centralized gateways.

Chapter 7

Evaluation

7.1 Lorem Ipsum

Identification within distributed networks: (*Decentralized identifiers (DIDs) v1.0*, 2019).

Chapter 8

Future Prospects

8.1 Lorem Ipsum

Future Prospects and outlook.

Chapter 9

Conclusion

9.1 Thesis summary

In summary, pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nunc eleifend, ex a luctus porttitor, felis ex suscipit tellus, ut sollicitudin sapien purus in libero. Nulla blandit eget urna vel tempus. Praesent fringilla dui sapien, sit amet egestas leo sollicitudin at.

9.2 Future work

There are several potential directions for extending this thesis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam gravida ipsum at tempor tincidunt. Aliquam ligula nisl, blandit et dui eu, eleifend tempus nibh. Nullam eleifend sapien eget ante hendrerit commodo. Pellentesque pharetra erat sit amet dapibus scelerisque.

Vestibulum suscipit tellus risus, faucibus vulputate orci lobortis eget. Nunc varius sem nisi. Nunc tempor magna sapien, euismod blandit elit pharetra sed. In dapibus magna convallis lectus sodales, a consequat sem euismod. Curabitur in interdum purus. Integer ultrices laoreet aliquet. Nulla vel dapibus urna. Nunc efficitur erat ac nisi auctor sodales.

References

Chawathe, Y., Ratnasamy, S., Breslau, L., Lanham, N., & Shenker, S. (2003). Making gnutella-like p2p systems scalable. *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications - SIGCOMM '03*, 407. https://doi.org/10.1145/863955.864000

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

Decentralized identifiers (DIDs) v1.0. (2019, December 9). https://www.w3.org/TR/2019/WD-did-core-20191209/

Guha, S. (2015). *An experimental study of the skype peer-to-peer VoIP system.* https://hdl.handle.net/1813/5711

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., & Swick, R. R. (2002). Annotea: An open RDF infrastructure for shared web annotations. *Computer Networks*, *39*(5), 589–608. https://doi.org/10.1016/S1389-1286(02)00220-7

Kirschenbaum, M. G. (2016). *Track changes: A literary history of word processing.* The Belknap Press of Harvard University Press.

Marshall, C. C. (1997). Annotation: From paper books to the digital library. *Proceedings of the Second ACM International Conference on Digital Libraries - DL '97*, 131–140. https://doi.org/10.1145/263690.263806

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia : Links, Objects, Time and Space—Structure in Hypermedia Systems: Links, Objects, Time and Space—Structure in Hypermedia Systems*, 40–49. https://doi.org/10.1145/276627.276632

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013). Designing the w3c open annotation data model. *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, 366–375. https://doi.org/10.1145/2464464.2464474

Siemens, L. (2015). "More hands" means "more ideas": Collaboration in the humanities. *Humanities*, *4*(3), 353–368. https://doi.org/10.3390/h4030353

Simon, R., Barker, E., Isaksen, L., & Cañamares, P. de S. (2015). Linking early geospatial documents, one place at a time: Annotation of geographic documents with recogito. *E-Perimetron*, *10*(2), 49–59. http://oro.open.ac.uk/43613/

Simon, R., Barker, E., Isaksen, L., & De Soto CaÑamares, P. (2017). Linked data annotation without the pointy brackets: Introducing recogito 2. *Journal of Map & Geography Libraries*, *13*(1), 111–132. https://doi.org/10.1080/15420353.2017.1307303

Terras, M. M., Nyhan, J., & Vanhoutte, E. (2016). *Defining digital humanities: A reader.* Routledge. https://www.taylorfrancis.com/books/e/9781315576251

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S.,

Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18