Findable and reusable?

Summary / Samenvatting

Summary

The reuse of research data is heralded as having the potential to increase effectiveness, productivity, and reproducibility in science. Stakeholders from funders to systems designers work to create policies, repositories and tools to support and encourage opening and sharing data. An assumption underlying this work is that data will be reused if they are shared. Another assumption predicating reuse is that data will be discovered by researchers, although relatively little empirical work exists to support this assumption.

This dissertation presents empirical research about the practices involved in discovering data, examining what researchers across disciplines are actually doing as they *find, make sense of and use data* which they have not created. It brings together perspectives from information science (IS), computer science (CS) and science and technology studies (STS) to examine these practices with the aim of informing and intervening into the design of search solutions for research data.

As detailed in Chapter 2, user-centered models of information seeking, common in IS and CS, provide an overarching theoretical framework for this research, while concepts from STS are used to engage with these models and explore the sociotechnical nature of data discovery. A range of qualitative and quantitative methods, including semi-structured interviews, observations, a bibliometric analysis and a large-scale survey, are used throughout the thesis to draw out the complexity of practice while also identifying patterns useful for the design of data search systems.

The first empirical study, presented in Chapter 3, analytically reviews the documented data discovery practices of users of observational data across five broad disciplinary groups: astronomy, earth & environmental sciences, biomedicine, field archeology and the social sciences. The review seeks to identify commonalities in how researchers in these disciplines discover and evaluate observational data for reuse, but it also documents the diversity of data needs and practices which exist. This chapter highlights the absence of particular communities in the reviewed literature and concludes that a theoretical framework based on information retrieval alone is insufficient for deeply understanding practices of data discovery.

Chapter 4 brings together bibliometrics and semi-structured interviews to further explore data discovery practices. A bibliometric analysis of the literature corpus collected in the third chapter reveals the technical bias and distributed nature of the discourse surrounding data

search. Interviews with data seekers connect two different perspectives - that of data seekers and systems designers, whose input informed the development of the interview protocol. The chapter calls for a broader understanding of both the individuals seeking and using data, as well as the data needed in research. It also recognizes the liminality of practice, finding that data discovery spans the thresholds of other practices, occurring, for example during the course of data sharing and data management. It explores how adopting a contextual, sociotechnical perspective can help to understand user practices and behavior and ultimately help to improve the design of data discovery systems.

The study presented in Chapter 5 employs a broader approach - a globally distributed multidisciplinary survey, with nearly 1700 respondents - to explore data discovery practices at a larger scale. The survey questionnaire was designed to probe the role of social interactions in seeking data and to explore relations to other practices, such as searching for academic literature. An initial quantitative analysis and a more extensive exploration of the qualitative data was used to consider commonalities in practices, as well as to examine differences, looking at how data needs and search practices vary not only by disciplinary domain but also by types of data uses. The chapter suggests how data communities can be conceptualized and proposes practical applications of the findings for designers of data discovery systems and repositories.

Chapter 6, which consists of two parts, represents my first-person experience with sharing and preparing data for reuse. The main body of the chapter, published as a data descriptor paper, details the openly-available data collected from the survey described in Chapter 5. This is followed by a reflection piece, in which I consider the process of sharing and describing my data; the affordances and duplication problems data papers present; and the role of data papers in the broader landscape of scholarly communication.

The final empirical chapter, Chapter 7, focuses on a particular part of the data discovery process: data-centric sensemaking. This study, the result of an intense international collaboration, combines a think-aloud task, a screen recording analysis and in-depth interviews with researchers, as they summarized and interacted with both familiar and unfamiliar data. The chapter identifies and details common patterns of data-centric sensemaking across three clusters of activities: *inspecting* data, *engaging* with content, and *placing* data within broader contexts. It also explores the role of contextual information and collaborative practices in understanding data and proposes specific design recommendations to facilitate sensemaking and subsequent data reuse.

This dissertation presents evidence of the diversity and multiplicity of data needs, discovery strategies and data uses existing both across and within disciplines. It finds that discovery practices are interwoven with other data and (re)search practices, i.e. searching for academic literature. Evidence is found for the varying ways which researchers make use of social connections - with data authors, peer networks, collaborators and broader communities - to locate and understand data. This thesis also emphasizes that discovery practices are shaped by data characteristics, as well as the hidden work involved in creating data and making them portable, i.e. processes of data description, sharing and standardization.

As a whole, this research contributes to the broader discourse on data discovery, sharing and reuse, as well as making recommendations (summarized in the final chapter) for the development of technologies which take into account the actual practices of researchers. Ultimately, this dissertation concludes that finding and reusing data are not reducible to performing keyword searches online or pressing a button to download data, demonstrating that data discovery and reuse are deeply sociotechnical practices, reliant on dynamic relationships between people, technologies, materials and policy.

Samenvatting

Vindbaar en herbruikbaar? Onderzoeksdata ontdekken in de praktijk

Hergebruik van onderzoeksdata wordt toegejuicht als mogelijkheid om de effectiviteit, productiviteit en reproduceerbaarheid van wetenschappelijk onderzoek te vergroten. Belanghebbenden, van financiers tot systeemontwerpers, werken samen aan regels, repositories en tools die de toegankelijkheid en beschikbaarheid van data ondersteunen en bevorderen. Een onderliggende aanname is daarbij dat gegevens worden hergebruikt als ze worden gedeeld. Een andere aanname ten gunste van hergebruik is dat de gegevens ook gevonden worden door onderzoekers, hoewel hiervoor relatief weinig empirisch bewijs bestaat.

In deze dissertatie wordt empirisch onderzoek gepresenteerd over de praktijken met betrekking tot de ontdekking van data, waarbij wordt bestudeerd wat onderzoekers in verschillende disciplines daadwerkelijk doen wanneer ze *data vinden, interpreteren en gebruiken* die ze niet zelf hebben gemaakt. De dissertatie combineert perspectieven uit de ‘information science’ (IS), ‘computer science’ (CS) en ‘science and technology studies’ (STS) om deze praktijken te onderzoeken ten behoeve van het ontwerp van zoeksystemen voor onderzoeksdata.

Zoals wordt uiteengezet in hoofdstuk 2, bieden gebruikersgerichte informatiezoekmodellen, die veel voorkomen in IS en CS, een overkoepelend theoretisch kader voor dit onderzoek, terwijl begrippen uit STS worden gebruikt om met deze modellen aan de slag te gaan en de sociaal-technische aard van data-ontdekking te verkennen. Een scala van kwalitatieve en kwantitatieve methoden, inclusief semigestructureerde interviews, observaties, een bibliometrische analyse en een grootschalige enquête, wordt in dit promotieonderzoek ingezet om de complexiteit van de praktijk aan het licht te brengen en ook om patronen aan te wijzen die bruikbaar zijn voor het ontwerp van datazoeksystemen.

In de eerste empirische studie, gepresenteerd in hoofdstuk 3, worden de gedocumenteerde data-ontdekkingspraktijken van gebruikers van observatiedata in vijf brede disciplinegroepen geanalyseerd: astronomie, aard- en milieuwetenschappen, biogeneeskunde, veldarcheologie en sociale wetenschappen. In de analyse wordt gezocht naar gemeenschappelijkheden in de manier waarop onderzoekers in deze disciplines

observatiedata ontdekken en beoordelen met het oog op hergebruik, maar wordt ook de diversiteit van de bestaande databehoeften en -praktijken in kaart gebracht. Het hoofdstuk belicht de afwezigheid van bepaalde gemeenschappen in de onderzochte literatuur en concludeert dat een theoretisch kader dat uitsluitend gebaseerd is op informatieontsluiting ontoereikend is voor een diep inzicht in de praktijk van de data-ontdekking.

In hoofdstuk 4 worden bibliometrische data gecombineerd met semigestructureerde interviews om nader onderzoek te doen naar de praktijk van het vinden van data. Een bibliometrische analyse van de in het derde hoofdstuk verzamelde literatuur brengt de technische bias en het gedistribueerde karakter van het discours rond datazoekopdrachten aan het licht. Interviews met datazoekers leggen een link tussen twee verschillende perspectieven, dat van de zoekers en dat van de systeemontwerpers wier input heeft bijgedragen aan de ontwikkeling van het interviewprotocol. In het hoofdstuk wordt gepleit voor een breder inzicht, niet alleen in de mensen die data zoeken en gebruiken maar ook in de benodigde data zelf. Ook wordt de liminaliteit van de praktijk erkend, in die zin dat het zoeken van data de drempels naar andere praktijken overschrijdt, en bijvoorbeeld plaatsvindt tijdens datasharing en datamanagement. Verder wordt onderzocht hoe de hantering van een contextueel, sociaal-technisch perspectief tot beter begrip kan leiden voor de praktijken en het gedrag van gebruikers en uiteindelijk kan bijdragen aan een beter ontwerp van datazoeksystemen.

De in hoofdstuk 5 beschreven studie kent een bredere benadering en betreft een mondiaal verspreide multidisciplinaire enquête met bijna 1700 respondenten, bedoeld om datazoekpraktijken op grotere schaal te onderzoeken. De vragenlijst voor deze enquête is opgesteld om na te gaan welke rol sociale interacties spelen bij het zoeken van data en het verband te onderzoeken met andere praktijken, zoals het zoeken naar academische literatuur. Met een initiële kwantitatieve analyse en een meer uitgebreide verkenning van de kwalitatieve data is gekeken naar gemeenschappelijke elementen in de verschillende praktijken, maar zijn ook de verschillen onderzocht, bijvoorbeeld door te kijken hoe databehoeften en zoekpraktijken niet alleen per vakdomein maar ook per type datagebruik variëren. In het hoofdstuk worden suggesties gedaan voor de wijze waarop datagemeenschappen kunnen worden geconceptualiseerd, en worden praktische toepassingen van de uitkomsten voorgesteld voor ontwerpers van data-ontdekkingssystemen en repositories.

In hoofdstuk 6, dat uit twee delen bestaat, beschrijf ik mijn eigen ervaringen met het delen en prepareren van data voor hergebruik. De bulk van het hoofdstuk, eerder gepubliceerd als *data descriptor paper*, beschrijft de openbaar beschikbare data die zijn verzameld op basis van de in hoofdstuk 5 beschreven enquête. Dit deel wordt gevolgd door een reflectief deel, waarin ik inga op het proces van delen en beschrijven van mijn data, de perceptie- en duplicatieproblemen die datapapers opleveren, en de rol van datapapers in het bredere landschap van de wetenschappelijke communicatie.

Het laatste empirische hoofdstuk, hoofdstuk 7, is gericht op een bepaald onderdeel van het proces van data-ontdekking: datacentrische betekenisgeving. In deze studie, het resultaat van intensieve internationale samenwerking, worden een aantal elementen gecombineerd: een opdracht waarbij hardop gedacht wordt, een analyse van schermopnamen en diepte-interviews met onderzoekers terwijl zij bekende en onbekende data samenvatten en gebruikten. In het hoofdstuk worden veelvoorkomende patronen van datacentrische betekenisgeving ontdekt en beschreven voor drie clusters van activiteiten: *inspecteren* van data, *aan de slag gaan* met content en *plaatsen* van data in bredere contexten. Ook wordt onderzocht wat de rol is van contextuele informatie en samenwerkingspraktijken bij het begrijpen van data, en worden specifieke ontwerpaanbevelingen gedaan voor de bevordering van betekenisgeving en daaropvolgend hergebruik van data.

Deze dissertatie levert bewijzen voor de bestaande diversiteit en multipliciteit van databehoeften, ontdekkingsstrategieën en datagebruik tussen en binnen disciplines. Het blijkt dat ontdekkingspraktijken verweven zijn met andere data- en (onder)zoekpraktijken, namelijk bij het zoeken naar academische literatuur. Er worden bewijzen gevonden voor de verschillende manieren waarop onderzoekers gebruikmaken van sociale verbindingen - met auteurs, peer-netwerken, samenwerkingspartners en grotere gemeenschappen - om data te lokaliseren en te interpreteren. In deze dissertatie wordt tevens benadrukt dat ontdekkingspraktijken worden gevormd door datakenmerken, en wordt aandacht besteed aan het onzichtbare werk dat verricht wordt om data te creëren en overdraagbaar te maken, dat wil zeggen de beschrijving, het delen en de standaardisatie van data.

Dit onderzoek als geheel draagt bij aan het bredere discours over het ontdekken, delen en hergebruiken van data, en heeft geleid tot aanbevelingen (samengevat in het laatste hoofdstuk) voor de ontwikkeling van technologieën die rekening houden met reële onderzoekspraktijken. Uiteindelijk is de conclusie van deze dissertatie dat het vinden en hergebruiken van data niet kan worden gereduceerd tot het uitvoeren van een online-zoekopdracht of een druk op de knop om data te downloaden, omdat is aangetoond dat data

vinden en data hergebruiken zeer sociaal-technische praktijken zijn die afhangen van dynamische relaties tussen mensen, technologieën, materialen en beleidsregels.