Supporting Relation-Finding in Neuroscientific Text Collections using Augmented Reality: A Design Exploration

A Master's thesis by Ivar Troost

Supervised by
Lynda Hardman & Wolfgang Hürst

June 12th, 2020 // ICA-4004477
Utrecht University

Maintaining an overview of publications in the neuroscientific field is challenging, especially with an eye to finding relations at scale (between, e.g., brain regions and diseases) – both those well-studied and nascent. To support neuroscientists in this challenge, we used a design-based research approach to investigate whether Augmented Reality could serve as a platform to make automated methods more accessible and integrated into current work practices. Building on insights from Text and Immersive Analytics, as well as two prior user studies, we identified information and design requirements (e.g., "highlight, not hide" and "augment, not replace"), which we embodied in a system design and implementation focussed on the analysis of co-occurrences in text collections. We evaluated our system using a scenario-based video survey with a diverse sample of neuroscientists and other domain experts, focusing on the quality of our design choices and participants' willingness to adopt such an AR system in their regular literature review practices. The AR-tailored epistemic and representation design of our system were generally perceived as suitable for performing complex analytics. We therefore see opportunities in pushing our generalisable interaction paradigm further in augmenting intellectual activities. We also discuss several fundamental issues with our chosen 3D visualisations, making steps towards addressing the question whether AR a suitable medium for relation-finding in document collections.

Scientific progress relies on the efficient assimilation of existing knowledge in order to choose the most promising way forward and to minimize re-invention.
– Tshitoyan et al. (2019)

The more papers are published, the harder it becomes for neuroscientists to maintain an overview. While literature reviews serve to defragment contributions, manual review is costly and – even when performed rigorously – researchers run the risk of missing important perspectives not identified at some part of the review process (Görg et al., 2010). Auer et al. argue that this bottleneck is inherent to *document-centric publishing*, and advocate for a move towards "expressing and representing information as structured, interlinked and semantically rich knowledge graphs" (2018, p. 1).

Talking with neuroscientists at the Institute of Automation of the Chinese Academy of Sciences, we found that one of the main shortcomings of manual literature exploration lies in performing complex relation-finding (Cunqing Huangfu, personal communication, June 12, 2019[1]; July 3, 2019). Without automated tools, it would be hopeless to find out which brain region is most often referenced when discussing, e.g., depression – indicating wide consensus on this relationship. Likewise, it would not be possible to find out which brain regions are mentioned only seldom – which could offer fruitful grounds for further investigation.

On both the computational and visualisation side (cf. Gopalakrishnan et al., 2019; Federico et al., 2017), computer scientists have addressed the need for a distant reading approach to academic literature. We are not aware that any

Acknowledgements: The thesis you are holding could not have been written without the significant effort of my two supervisors. Thank you, Lynda, for always having made time to discuss conundrums, offer feedback, and even narrate an entire video tutorial series. Your attention to detail has certainly upped the standard in my writing (though any remaining issues are fully mine). Wolfgang, thank you for having kept me focused during this complex design-based research project and making me better appreciate the boundaries between engineering and research. I should also thank the other members of the DatAR research team: Ghazaleh Tanhaei (as its captain, initiator, and a close collaborator from the start), Juliana Montaña Melgarejo (working on adding brains to our system), Ben van Klingeren (exploring AR–desktop hybrids), as well as Finn van der Heide (who contributed significantly to the first design cycle). Without their critical feedback and our weekly talks, this journey would surely have been more lonely. I would also like to thank Cunqing Huangfu for helping my understand the problem domain better and for his attentiveness in resolving any BAG-related data problems I posed. Tessa Heeroma and Anna van Harmelen also contributed to a crucial piece of the puzzle by linking the different external data sources. Finally, I would like to thank all participants who contributed their voices.

[1] With Ghazaleh Tanhaei

such system has yielded significant user adoption, however. More work is required to make tools directly accessible to neuroscientists (without the need for a supporting data analyst), and to integrate these in existing research workflows.

At the start of our design process, we resolved to investigate whether Augmented Reality (AR) could provide a suitable platform to complement current literature exploration practices – in particular by making automated methods more accessible. Recent work in Immersive Analytics (IA; Marriott, Schreiber, et al., 2018) highlight the intuitiveness of natural user interfaces, as well as the immersive properties of IA visualisations and the opportunity to render three-dimensional data binocularly (e.g., brain visualisations). Moreover, conducting literature explorations over longer periods of time could benefit from persistently placed virtual objects in physical space (c.f. *Method of Loci*).

Our study uniquely contributes to the design space where Text and Immersive Analytics overlap. We present the design process and evaluation of a novel analytics system: DatAR (*DATa exploration in Augmented Reality*). Our system seeks to harness the visuospatial representations that Augmented Reality enable to provide access to neuroscientific knowledge graphs generated by automated methods. The end goal is to support neuroscientists in performing complex relation-finding tasks through human-in-the-loop analytics. We evaluated our system using a scenario-based video survey, focusing on the quality of our design choices and participants' willingness to adopt the system in their regular practice.

Related Work

DatAR exists at the crossroads of two research agendas. We focus on the challenge of Text Analytics and apply solutions offered by Immersive Analytics. In this section we cover prior work in these areas.

Text Analytics

There is already fertile ground in computer-supported relation-finding in large document collections, especially where it regards scholarly literature. We will highlight two academic streams in this subsection: literature-based discovery (with a focus on algorithms) and text visualisation (with a focus on visuospatial representations).

Literature-Based Discovery

Unsupervised learning methods can be used to extract semantic relationships from published literature at scale. Literature-based discovery is a research strategy that uses various statistical and machine-learning techniques "to exploit already known scientific knowledge to generate hitherto unknown but meaningful connections" (Gopalakrishnan et al., 2019, p. 2). Using this information, scholars can make better decisions as to which hypotheses to pursue next. Lander (2010) stresses that the scientific relevance of such findings is not in simply stating new links, but rather in *understanding why* these links exists. In other words, it requires developing *scientific models* through careful human-in-the-loop analytics. This requires both domain knowledge as well as know-how in using the appropriate computational tools.

Gopalakrishnan et al. (2019) distinguished several categories of literature-based discovery in their survey in the biomedical domain. The field started off by using *co-occurrence approaches* (c.f., Swanson, 1986). The starting point here is an insight from linguistics, the *distributional hypothesis*: if two concepts are repeatedly mentioned in the same analysis unit (e.g., in a paper, abstract, or sentence), they are semantically associated. *Semantic relation-based approaches* additionally attempt to capture the meaning behind the association of concepts. For example, a particular drug could have a negative or positive effect. *Graph-based approaches* use graph theory to address scalability of analyses to allow more sophisticated queries on the data. This is especially relevant when trying to identify associations that are still in early stages of discovery. In this latter category, we find the work by Gramatica et al. (2014), whose approach is comparable to the one taken by the providers of our primary data set (the *Brain Association Graph*; see the DatAR System Design section). Gramatica et al. (2014) used knowledge graph analysis to find high-potential drug–disease combinations that were not yet studied in-depth. They constructed their knowledge graph by mining concept co-occurrences in PubMed paper abstracts. These concepts were derived from a predetermined dictionary. Subsequently, they used graph algorithms (e.g., random walk distance) to find the shortest paths between drugs and diseases. Based on their results, they determined several new uses for existing drugs.

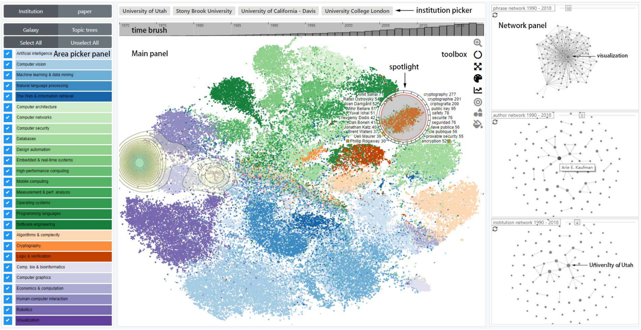
Text Visualisation

Attempts have been made at representing documents by using, e.g., graphs, point clouds (in 2D and 3D), contour maps, and geometrical grids (Li et al., 2019). Federico et al. (2017) offer a useful taxonomy of text visualisation approaches based on the data researchers use: Text, Citations, Authors, and Metadata. The DatAR system currently restricts itself to Text: research paper abstracts.

DatAR uses a topic model visualisation to spatially organise neuroscientific document collections in 3D space. In a regular topic model, an algorithm such as LDA is used to generate n topics based on recurrent word use in a set of given documents. Each document–topic pair is assigned a probability (such that for each document the accumulative probability is 1). This yields an n-dimensional space, which can be visualised by collapsing it to 2–3 dimensions (using a dimension reduction algorithm, e.g., t-SNE). The distance between documents represents their semantic similarity: the

Figure 1

The Galex interface developed by Li et al. (2019). Using text-mining and thoughtful interactive affordances, the authors made the evolution of scientific areas in computation visible. © 2019 IEEE



closer they are, the more similar (based on the topics assigned). This approach offers an effective means of analysing large data sets through visuospatial distant reading. A good example of this is offered by Li et al. (2019), who showcase a multifaceted user interface that allows users to explore subtopics and papers in the computer sciences (see Figure 1). Their interface uses data generated through a combination of thesauri (containing predetermined concepts) and topic modelling with dimension reduction (using doc2vec, LDA, and t-SNE algorithms). The final visualisation offers an effective overview of a large document collection, additionally allowing interactive identification and further inspection of document clusters.

Immersive Analytics

DatAR differentiates itself from previous work in Text Analytics by focusing on the analytics experience *in Augmented Reality*. While Visual Analytics has become a mainstay in human-in-the-loop data analysis, Immersive Analytics (IA) attempts to move analytics tools from the 2D screen into our environment (both in Virtual and Augmented Reality; for a survey of the field, see Fonnet and Prié, 2019). The aim is to design "engaging, embodied analysis tools to support data understanding and decision making [and] liberate these activities from the office desktop" (Marriott, Chen, et al., 2018, p.14–15).

Immersive Analytics research has its roots in the use of large curved wall displays (CAVEs). However, findings by Cordeil, Dwyer, et al. (2017) suggest that modern head-mounted displays (HMDs) have caught up by being both more cost-effective and allowing faster analysis without sacrificing accuracy. Moreover, HMDs output binocular imagery, aiding depth perception. Given that affordance, a key consideration is whether to map data to 2D or 3D space. Any use of the third dimension has long received strong scepticism within the information visualisation community due to the added visual complexity (Marriott, Chen, et al., 2018), but there is now a renewed interest in critically re-assessing *binocular* 3D visualisations. As every reduction of an *n*-dimensional space (such as the output of a topic model) translates to data loss (Gracia et al., 2016), our project intends to re-evaluate the merits of 3D information visualisation.

Finally, there are also several academic toolkits available in the area of IA, such as DXR (Sicat et al., 2019) and IATK (Cordeil, Dwyer, et al., 2017), based on such works as ImAxes (Cordeil, Cunningham, et al., 2017). However, these frameworks were developed with quantitative data in mind. While the database that DatAR uses does contain some numerical data, its most important features are its relationship topology and the text it contains. Graphs and text require different visualisation strategies, which is why we have opted to develop our system from the ground up with this in mind, rather than adopt and work around a framework.

Methods

Hevner et al. (2004) argue that progress in Information Systems requires two "complementary but distinct paradigms" (p. 3): behavioural science – which attempts to explain or predict phenomena and aims to generate *understanding* – and design science – which attempts to change phenomena and aims to generate *utility*. As our main research aim was to deliver an immersive analytics environment that would offer a glimpse of a speculative future of processing academic literature, we opted for a design research strategy[2]. Using the language of Gregor and Hevner (2013), our angle of approach was *exaptation*: to take a known – albeit experimental – solution (Immersive Analytics) and apply it to a new problem (Text Analytics). We should emphasise that we report on one of the first design cycles of the project and that this study's main purpose is to set a course for further theory development in the next design iterations.

Given the large influence it has had on our work, it would be prudent to explicate our epistemological assumptions. We have taken a constructivist approach, taking notes from Papert's (1980) constructionism – in particular agreeing that we need environments where knowledge can be played with, and where externalised cognitive representations play a key role in reaching understanding. This view is closely related to that of distributed cognition, popularised in Information Visualisation (Liu et al., 2019; also see Rogers, 2012). Distributed cognition views cognition as "embodied, enculturated, situated in local interactions, and distributed or stretched across humans and artifacts" (Liu et al., 2019, p. 1174). That means that the object of study is not the user (who has internal representations) or the artifact (which contains external represen-

[2] A more thorough survey of design research approaches can be found in Appendix J

tation) in isolation – it is the cognitively coupled system that they represent together, and how coordination between parts of this system allows for successful sense- and decision-making. Taking the coupled cognitive system as our unit of analysis precludes a divide-and-conquer research strategy, i.e., selectively picking one dimension or aspect without considering their context (Liu et al., 2019).

Due to the novelty of our approach, we decided a tight design–evaluate loop was necessary in the early stages of the project. Without precedent in many aspects of our design space, we had to base features on the best available information and evaluate their soundness empirically. For this purpose, we organised two pilot user studies. The first pilot study was organised as a workshop and held with six experts in data representation and visualisation, with the goal of gathering implementation-oriented feedback. We used the framework by Kerzner et al. (2019) to guide the organisation of this session and to elicit feedback from our participants. After the workshop segment, participants had the opportunity to try the current prototype of the system and offer feedback. This is also the period that our research group presented a DatAR research proposal at AIVR 2019 (Tanhaei et al., 2019).

The second pilot study was scenario-based and held with eight bachelor-level neuroscientists, with the goal of gathering domain-oriented feedback. In a lab setting, participants performed a researcher-guided data exploration scenario. Afterwards, we conducted semi-structured interviews in which current review practices were discussed, and in which the system was evaluated on meaningfulness, visualisation and representation quality, and navigation and interaction quality. We distilled and integrated the critiques from both pilots in the third design cycle, which we present here. We concluded the present cycle with a video survey among academic peers. We will return to this in detail in the Evaluation Study section.

Finally, we should comment on how we understand and safeguard the scientific rigour of our work. We follow the interpretivist framework by Meyer and Dykes (2019), who formulated criteria a work can be judged by after the design process has completed. The authors expect a good design study to be: (1) informed by already existing knowledge, (2) reflexive of the researchers' own role in the study, (3) abundant in having considered and tried many possibilities, and using rich descriptions to convey information, (4) plausible in making knowledge claims that are evidence-based, context-aware and persuasive, (5) resonant by being transferable and evocative, and (6) transparent in being particular enough about reporting. We adopt these values as criteria to meet, and will return to these in the Discussion section.

DatAR System Design

In the Data section we discuss what Sedig et al. (2012) dubbed the *information* and *computing spaces* of the design product: where raw sources of information live, and where this information is encoded, stored, mined and transformed. Given that input, the task of relation-finding, and results from our pilot studies, we move on to the Design Requirements: the needs and principles we aimed to embody in DatAR. In the final two sections we discuss Design and Implementation of these requirements, in terms of information representation and user interaction.

Data

The main data set we use is the Brain Association Graph (BAG), developed as part of the Linked Brain Data platform[3]. This is a triple store[4] that contains co-occurrences of known concepts[5]. The database was created by mining PubMed abstracts for co-occurring concepts of different types within sentences (for example, hippocampus – a brain region – and Alzheimer's – a disease)[6]. The BAG currently contains well over a 100,000 sentences containing co-occurrences of a region and a disease[7]. Statistics were calculated on each concept pair (e.g., hippocampus–Alzheimer's) to determine the total count of co-occurrences, and the two-way probabilities of concepts being in the same sentence. The Pubmed ID of the paper of each sentence is kept intact during this process, allowing for fetching additional metadata (such as date of publication, authors and venue).

The BAG web portal displays the stored triple data in table form by category pair (e.g., region–disease). This visualisation is suboptimal, however, given the underlying graph-based nature of the dataset and a lack of filtering, sorting, and aggregating functionalities. There is also an open endpoint that allows users to specify queries using the SPARQL syntax[8]. While SPARQL enables powerful query capabilities, using such a query language requires training. DatAR

[3] http://www.linked-brain-data.org

[4] A type of graph database that stores all its data as Subject-Predicate-Object triples; also known as RDF. An example of such a triple is: *lbd:amygdala rdf:type lbd:region*. Note that each of these items start with a namespace, which is shorthand for a full URI. For a primer, you may refer to Antoniou et al., 2012.

[5] We use the word *concept* to refer to resources in linked open data repositories as well as the entities that they refer to. E.g., lbd:amygdala (the signifier) and Amygdala (the signified).

[6] Keep in mind that we are not looking at actual medical relationships here – we are looking at how scholars describe these relationships in their publications.

[7] We restrict ourselves to regions and diseases in our present work, but neurons, proteins, genes, and neurotransmitters are also available.

[8] For details on the SPARQL query language, you may refer to W3C (2013)

aims to abstract away from these infrastructure details for the user (while still making them aware of the provenance of the data).

While the BAG serves as the central data repository, we have (manually) connected its concepts to other linked data repositories to extend the analytic possibilities. We use MeSH[9] and Wikidata[10] to offer additional descriptions of individual brain regions and diseases. We have also linked brain regions in the BAG to the Scalable Brain Atlas[11] – allowing volumetric localisation of regions using BAG concepts. For more details on our data mappings and storage infrastructure, see Appendix C.

Finally, we have received a reduced-dimension topic model of brain diseases from Cunqing Huangfu, custodian of the BAG. He used all sentences in the BAG that included at least one region as source data. All sentences describing a disease were combined into single documents, which were then processed with LDA (topic modelling) and T-SNE (dimension reducation) algorithms. This resulted in a three-dimensional coordinate space, in which the distance between diseases represents their semantic similarity (the closer they are, the more similar). This data was used as input to the *Topic Model Visualisation Widget*, described in more detail in the Representational Design section.

Design Requirements

Our top-level goal is to support neuroscientists in relation-finding. Using related work and an informal task analysis, we stipulated several design requirements and guiding principles to reach this goal. Some of these were present from the very start of our process, and others emerged through interaction with early prototypes of the system as well as the pilot studies. For provenance of design requirements, please refer to Appendix I.

Supporting Open and Closed Discovery

In literature-based discovery, the distinction is made between open and closed discovery tasks (Gopalakrishnan et al., 2019). Whereas closed discovery tasks focus on better understanding the (direct and indirect) relationships between two predetermined concepts, open discovery attempts to identify and study all pertinent relationships originating from a single concept. As our goal is to support *finding* relationships (rather than inspecting them) our system must support open discovery: Given any entity in the data set, the user must be able to see which other entities it is related to. Moreover, users should be able to find strong relations (indicating common knowledge) as well as weaker ones (which could offer opportunities for further research).

Highlight, Not Hide

A key feature of relation-finding is that the user does not know what they are looking for prior to their analysis. The user therefore has to able to situate themselves in the entirety of the data set. We concur with Woods et al. (2002) that *perceptual organisation* that preserves context is therefore a necessity: Users need to be assisted in highlighting subsets of the data to improve observability rather than be faced with a system that hides presumably irrelevant aspects. Highlighting based on perceptual organisation requires modelling the domain semantics (neurosciences), to create a more abstract *conceptual space* to navigate. In our case, we would need to distinguish between the different classes of concepts (regions, diseases, etc.) as well as their co-occurrences (including aggregate statistics). The BAG already offers this structure, challenging us to find suitable ways of organising this data visuospatially such that we utilise high-bandwidth perceptual channels without overwhelming users.

Augment, Not Replace

Our approach should distinguish itself by being fully additional to current (screen- and paper-based) literature exploration practices rather than replacing these. As most such analytics work takes place on the desktop, it would be beneficial to integrate with both physical elements (i.e., Situated Analytics; Thomas et al., 2018) as well as implement bridges to other devices.

Making Use of the Medium

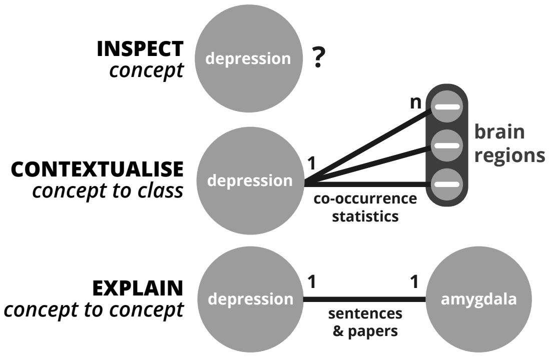
Augmented Reality lends itself to different interactions from traditional desktop environments. Our interface should therefore limit itself to the two affordances dependably supported by hand-tracking technology in AR HMDs: the hand as 3D cursor and grabbing. While keyboard (and other peripheral) support is arguably also part of AR's appeal, as compared to VR, we set as a goal to stick to the core interaction paradigm of AR where possible; any added peripheral would make the system less portable. In addition, we set out to put to use the infinite spatial canvas of AR by building a decentralised interface: All functionality should be local rather than global, and there ought not to be any singletons where data views and representations are concerned. This enables a form of subjunctive interface, "in which multiple queries can be explored simultaneously" (Bron, 2013, p. 4), which were shown to bolster breadth-first search behaviour in an explorative search task in media studies (Bron, 2013).

Additionally, we strove to make our system modular and internally consistent, with an eye on open-sourcing the code to allow other researchers to use DatAR as a toolkit for testing new visualisation approaches.

[9] https://id.nlm.nih.gov/mesh/
[10] https://www.wikidata.org/
[11] https://scalablebrainatlas.incf.org/

Figure 2

Epistemic design: Tasks supported in DatAR (Inspect, Contextualise, Explain). The concepts mentioned, depression and amygdala, are examples.



Design

Epistemic Design

Based on our review of literature-based discovery and our design requirements, we support three core tasks for relation-finding activities in document collections (Figure 2). Firstly, users need to be able to *inspect* lesser known concepts, retrieve their definitions, and find similar concepts. Secondly, they require the capability to *contextualise* a particular concept in regards to a *set* of other concepts. In our system, users can relate any disease to all brain regions, or any region with all diseases. Finally, users should be able to *explain* any given relation found by the system by requesting the sentences that attest to that relationship.

Representational Design

To enable the operations required to perform the tasks in the Epistemic Design, we decided to have concepts be visuospatially manifested as **Resource Spheres (RSs** ; see Figure 3 for a depiction, also for other representations in this section). RSs contain a URI, a user-facing label, and a class[12] – which are pulled from the BAG or other data sources. Users can grab RSs using a grabbing gesture and move them at will, fitting in with the AR interaction paradigm. RSs are used as input to Widgets (or can be output by them). **Widgets** are analysis tools that perform actions such as querying, data manipulation, visualisation or data export. A Widget can be standalone, requiring no further user input. The *Available Types Querier*, for example, renders all available concept types in the data storage (e.g., Disease, Region) as RSs without further instructions. Other Widgets have one or more **Receptacles**, in which an RS needs to be placed. For example, the *Resource Sphere Inspector* is a Widget that, after placing a RS in its Receptacle, pulls descriptions and closely matched concepts from Wikidata, MeSH and the Scaleable Brain project. These descriptions are then displayed to the user; the concepts are output as new RSs. This Widget addresses the Inspect task goal we set.

Some Widgets allow or require a **Dataflow** as input rather than a RS. Dataflows contain lists of concepts, and allow Widgets to communicate with each other[13]. Widgets can have a Dataflow **Outlet** (for Query Widgets), **Inlet** (for Visualisation and Export Widgets), or both (for Manipulation Widgets). An Outlet and Inlet can be connected by holding them together (where each Outlet can connect to multiple Inlets). An example of a Widget using Dataflows is the *Dataflow Inspector*, which renders incoming contents as a list. Dataflows allow chaining multiple Widgets, which automatically update if anything prior in the chain is modified.

Table 1

Textual overview of DatAR representations; legend for Figure 3.

#	Type	Representation
0	-	Resource Sphere
1	Q	Concepts of Class Querier
2	Q	Available Classes Querier
3	Q	Co-occurrence Querier
4	M	Min-Max Filter
5	V	Topic Model Visualisation
6	V	Brain Model Visualisation
7	V	Dataflow Inspector
8	V	Resource Sphere Inspector
9	E	Concept Pair Exporter

Note. Types used: Query (Q), Manipulation (M), Visualisation (V), and Export (E).

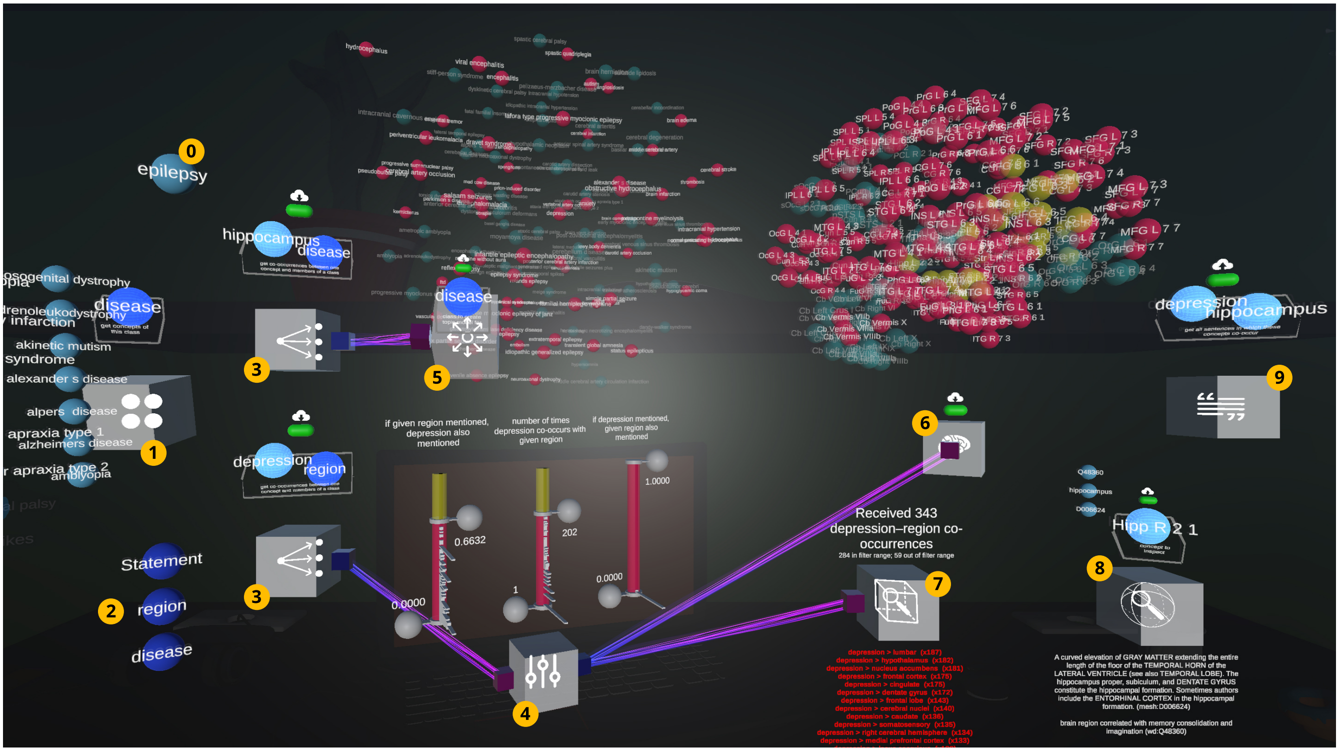
Our system takes into account the open-world assumption (i.e., our data set could contain concepts of any type, including unknown ones). Widgets are therefore responsible for specifying which data types are acceptable for it to process. For example, the *Co-Occurrence Querier* requires a concept of any type, and a class. It will subsequently try to find relations between the given concept and any items of the given class, and output these as a dataflow. The *Concept Pair Exporter* requires one concept of type Region and one of type Disease, and sends this information to a web-based companion application (see Figure 4). This web interface subsequently queries the BAG and PubMed to retrieve sentences (and their papers) containing both concepts – satisfying the Explain task goal we set. The *Min-Max Filter* supports Dataflows that pass a co-occurrence list; it then outputs a modified co-occurrence list based on the filter parameters.

[12] For example, lbd:amygdala as URI, "Amygdala" as label, and lbd:region as class.

[13] Dataflows as visual representations of data have been used in VR before, with promising results where it concerns ease-of-use (Ens et al., 2017).

Figure 3

Visual overview of DatAR representations. See Table 1 for a legend. Notice the two Dataflow constellations, starting at Query nodes (3) and ending at Visualisation nodes (5 & 6). In the top constellation, we see a Topic Model Visualisation (5) in which all diseases co-occurring with the hippocampus are highlighted in red; other diseases are displayed in a low-opacity turquoise. In the bottom constellation, the output generated by the Co-Occurrence Querier (3) is first manipulated by a Min-Max Filter (4), in which the most prevalent co-occurrences (in absolute count and relative importance to the region) have been filtered out. This filter then passes on the data to a Dataflow Inspector (7), which renders a list of contents, and the Brain Region Visualisation (6). The latter shows three filter states for depression–region co-occurrences: Turquoise RSs indicates a complete lack of any detected co-occurrence of depression and the given brain region in the literature; yellow indicates these do exist, but outside of the filter range; and red indicates co-occurrences within filter range. These colours update live as the user moves the filter sliders.



Received 343 depression–region co-occurrences
284 in filter range; 59 out of filter range

depression > lumbar (x187)
depression > hypothalamus (x182)
depression > nucleus accumbens (x181)
depression > frontal cortex (x175)
depression > cingulate (x175)
depression > dentate gyrus (x172)
depression > frontal lobe (x143)
depression > cerebral nuclei (x140)
depression > caudate (x136)
depression > somatosensory (x135)
depression > right cerebral hemisphere (x134)
depression > medial prefrontal cortex (x133)

A curved elevation of GRAY MATTER extending the entire length of the floor of the TEMPORAL HORN of the LATERAL VENTRICLE (see also TEMPORAL LOBE). The hippocampus proper, subiculum, and DENTATE GYRUS constitute the hippocampal formation. Sometimes authors include the ENTORHINAL CORTEX in the hippocampal formation. (mesh:D006624)

brain region correlated with memory consolidation and imagination (wd:Q48360)

A final core mechanic is highlighting. Each item in a Dataflow has a highlight flag, which can be read and/or modified by Widgets. There are two filter states: (1) out of filter range, and (2) in filter range. This information allows downstream visualisations to render their contents differentially.

Visualisation Design

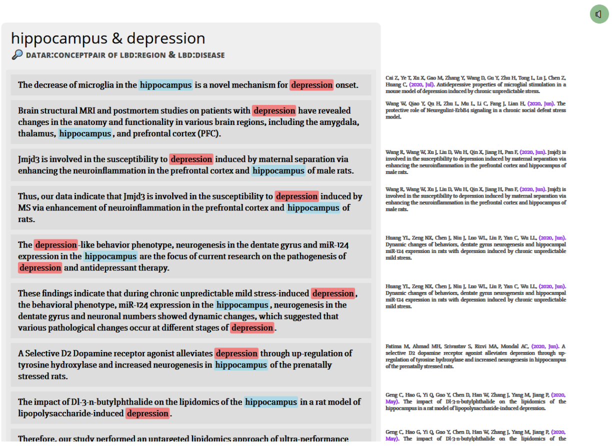
In this study, we implemented two main visualisations: a Topic Model and a Brain Region Visualisation (see the Data section for provenance). Conceptually, the *Topic Model Visualisation* takes in a class (RS) and returns a reduced topic model as a three-dimensional scatterplot of all concepts of that class. Each point is represented as a RS, which is replicated when a user tries to grab it. The Brain Region Visualisation behaves similarly, but instead displays the central points of brain regions in the Scalable Brain database.

While useful on their own, both visualisations become more useful when paired with other Widgets. When using Dataflows to connect the *Co-occurrences Querier*, concepts in both sets are highlighted. Adding a *Min-Max Filter* inbetween additionally allows differentiating concepts in and out of filter boundaries. This three-Widget set-up is currently the most powerful way of looking at the data within our system, satisfying the Contextualise task goal we set.

As with any complex design, we could never report all design decisions in a paper. For example, what the position of the user should be in relation to a 3D scatterplot. We have documented such considerations in our design and research documentation (see Appendices D and E), keeping a transparent and informed trace of decisions – following the design study criteria for rigour by Meyer and Dykes (2019).

Figure 4

DatAR Web Companion. This output was rendered after placing hippocampus and depression Resource Spheres into the Concept Pair Exporter. The rendered information includes the incoming concept pair, all sentences that contain both these concepts (where each occurrence is highlighted), and hyperlinked source papers.



hippocampus & depression

DATAR:CONCEPTPAIR OF LBD:REGION & LBD:DISEASE

The decrease of microglia in the hippocampus is a novel mechanism for depression onset.	Cai Z, Ye T, Xu X, Gao M, Zhang Y, Wang D, Gu Y, Zhu H, Tong L, Lu J, Chen Z, Huang C, (2020, Jul). Antidepressive properties of microglial stimulation in a mouse model of depression induced by chronic unpredictable stress.
Brain structural MRI and postmortem studies on patients with depression have revealed changes in the anatomy and functionality in various brain regions, including the amygdala, thalamus, hippocampus , and prefrontal cortex (PFC).	Wang W, Qiao Y, Qu H, Zhu L, Mu L, Li C, Fang J, Lian H, (2020, Jun). The protective role of Neuregulin1-ErbB4 signaling in a chronic social defeat stress model.
Jmjd3 is involved in the susceptibility to depression induced by maternal separation via enhancing the neuroinflammation in the prefrontal cortex and hippocampus of male rats.	Wang R, Wang W, Xu J, Liu D, Wu H, Qin X, Jiang H, Pan F, (2020, Jun). Jmjd3 is involved in the susceptibility to depression induced by maternal separation via enhancing the neuroinflammation in the prefrontal cortex and hippocampus of male rats.
Thus, our data indicate that Jmjd3 is involved in the susceptibility to depression induced by MS via enhancement of neuroinflammation in the prefrontal cortex and hippocampus of rats.	Wang R, Wang W, Xu J, Liu D, Wu H, Qin X, Jiang H, Pan F, (2020, Jun). Jmjd3 is involved in the susceptibility to depression induced by maternal separation via enhancing the neuroinflammation in the prefrontal cortex and hippocampus of male rats.
The depression-like behavior phenotype, neurogenesis in the dentate gyrus and miR-124 expression in the hippocampus are the focus of current research on the pathogenesis of depression and antidepressant therapy.	Huang YL, Zeng NX, Chen J, Niu J, Luo WL, Liu P, Yan C, Wu LL, (2020, Jun). Dynamic changes of behaviors, dentate gyrus neurogenesis and hippocampal miR-124 expression in rats with depression induced by chronic unpredictable mild stress.
These findings indicate that during chronic unpredictable mild stress-induced depression, the behavioral phenotype, miR-124 expression in the hippocampus , neurogenesis in the dentate gyrus and neuronal numbers showed dynamic changes, which suggested that various pathological changes occur at different stages of depression.	Huang YL, Zeng NX, Chen J, Niu J, Luo WL, Liu P, Yan C, Wu LL, (2020, Jun). Dynamic changes of behaviors, dentate gyrus neurogenesis and hippocampal miR-124 expression in rats with depression induced by chronic unpredictable mild stress.
A Selective D2 Dopamine receptor agonist alleviates depression through up-regulation of tyrosine hydroxylase and increased neurogenesis in hippocampus of the prenatally stressed rats.	Fatima M, Ahmad MH, Srivastav S, Rizvi MA, Mondal AC, (2020, Jun). A selective D2 dopamine receptor agonist alleviates depression through up-regulation of tyrosine hydroxylase and increased neurogenesis in hippocampus of the prenatally stressed rats.
The impact of Dl-3-n-butylphthalide on the lipidomics of the hippocampus in a rat model of lipopolysaccharide-induced depression.	Geng C, Hao G, Yi Q, Guo Y, Chen D, Han W, Zhang J, Yang M, Jiang P, (2020, May). The impact of Dl-3-n-butylphthalide on the lipidomics of the hippocampus in a rat model of lipopolysaccharide-induced depression.
Therefore, our study performed an untargeted lipidomics approach of ultra-performance	Geng C, Hao G, Yi Q, Guo Y, Chen D, Han W, Zhang J, Yang M, Jiang P, (2020, May). The impact of Dl-3-n-butylphthalide on the lipidomics of the

Implementation

We used Unity[14] (v2019.3.9f1) to build our main interface. Dataflows and Receptacles were developed using a reactive framework, UniRx[15] (v7.1.0), which allows for live-processing of data changes. The AR version of the system was built using the Meta SDK (v2.7.0.38) to support the Meta 2 HMD, optionally using Leap Motion for hand-tracking. The VR version of the system was built on SteamVR (v2.5; SDK v1.8.19)[16].

A companion web application was developed in Angular[17] (v8.2.14) to allow the main environment to send text-heavy content to a screen for easier reading than AR currently allows. To coordinate this communication, we used a RabbitMQ[18](v3.8.2) instance as a message broker.

The internal data structure in both the Unity and Web interfaces follows the guidelines of the JSON-LD standard[19], and is easily exported as such.

On a final note, we should emphasise that Widgets have been implemented modularly. This allows DatAR contributors to easily create new Widgets by combining existing building blocks (such as Receptacles, Inlets, and Outlets), and then adding custom data processing behaviour. For a full overview of these buildings blocks and JSON-LD examples, you may refer to the contributor's guide in Appendix B.

Evaluation Study

Objectives

Whereas in the previous two design cycles we emphasised Perceived Ease of Use (PEOU) and Perceived Usefulness (PU), in this third cycle we wanted to zoom out and (additionally) measure Attitude Toward Using (AT) and Behavioural Intention to Use (BI) DatAR-like systems among a larger and more diverse group of potential users. These are constructs from the Technology Acceptance Model (TAM), a theory from Information Systems that looks at which factors influence adoption of novel technologies (Rese et al., 2017; Chuttur, 2009; Davis, 1985). After all, it is important to look ahead before too many resources have been sunk into a design research project: Are researchers waiting for DatAR as a serious tool in their workflow? What are perceived opportunities in performing relation-finding in augmented reality, and which issues are critical?

Next to this, we wanted to evaluate whether participants experienced the design artefacts as we had hypothesised in the design rationale, and to what extent our epistemic, representation, and visualisation design were intuitive to grasp.

To these ends, we conducted a video survey[20]. While the disadvantage of this evaluation approach is that users do not get a hands-on experience with the system, it is more accessible and therefore allows a larger and more diverse group to participate. Given our objective and the TAM's focus on *perceptions* of technology, a high-fidelity mock-up scenario is a fitting alternative to user studies.

Methods

Participants

22 people participated in our video survey (10 women and 12 men, between the ages of 18–64, with a median age of 25–34). We used a convenience sample, drawing from colleagues of our institutes, a Facebook group of University College students, and participants who had joined our second design cycle's user study.

Seven of our participants were neuroscientists, working in the sub-fields of (cognitive) neurosciences, neurobiology, pharmacology, and psychiatry. Seven participants worked in computer science, data science/visualisation and statistics. Others were active in social sciences (4, of which 3 in educational sciences), digital humanities (1), and military sci-

[14] https://unity.com/

[15] https://github.com/neuecc/UniRx

[16] In the latest version of the DatAR system, v2020.1.1, only SteamVR is supported.

[17] https://angular.io/

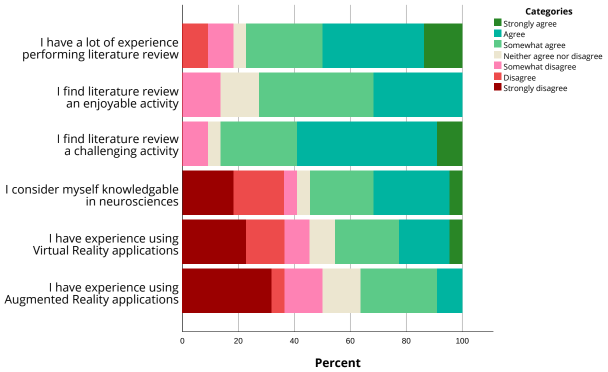
[18] https://rabbitmq.com/

[19] https://json-ld.org/

[20] We should note that the ideal second evaluation would have been to analyse sensemaking in users during their interaction with an improved system during a follow-up lab study. We had made preparations for this, but these plans were foiled by the COVID-19 outbreak. We took this as an opportunity to reach a larger pool of participants remotely.

Figure 5

Additional demographics data of our sample.



ence (1)[21]. Among our sample were participants of various statures, from bachelor's students to full professors[22]. We asked participants about their prior experiences with reviewing literature, the neuroscientific domain, and XR[23] platforms (see Figure 5). We report on correlations with these background factors in the Results section.

Materials & Protocol

To gather data, we administered a video survey (which took on average 37.23 minutes, *SD* = 13.98, to finish; see Appendix F for the scenario, videos, and questions). We introduced the survey to participants as a means of "exploring a speculative future in which researchers use augmented reality interfaces to support them in understanding complex relationships in their academic literature."

The system shown to users used VR to create a mock AR environment. This set-up was explained in the first video, in which we also showed an earlier AR version of the system. Given we already had to convert the system to VR to allow remote user studies during the COVID-19 period (this parallel evaluation will be reported elsewhere), we opted to use this version for the video survey as well. An added advantages was that this approach alleviated the immaturity of hand-tracking technology that we experienced in earlier design cycles. Our assumption in doing this research is that the quality of hand-tracking technology (and AR technology in general) will improve over the coming years, which is the timeline in which DatAR will become actually useful in practice (cf. Future Studies methods in HCI, Mankoff et al., 2013). Given our main interest is the willingness of participants to adopt technology in the future, using a higher-fidelity VR-based AR mock-up made more sense than showing an AR version with contemporary limitations. In a sense, our work could be considered a variant of the virtual field study (Mäkelä et al., 2020).

The survey consisted of four sections. Firstly, we inquired about the demographics (reported in the Participants section). Secondly, we asked participants to describe any specialised tools they uses for performing literature reviews – we did this to increase our understanding of how users currently tackle similar tasks. The third section contained seven tutorial videos and a mock-up scenario. According to Kyng, mock-ups scenarios aim to "simulate future use situations in order to allow end users to experience what it would be like to work with the system under development and thereby to draw on their tacit, non explicit knowledge and experience" (as cited in Rizzo and Bacigalupo, 2004, p. 6). In other words, they are a means to elicit a conversation with the user, which we tried to capture in our (asynchronous) survey by offering ample comment space below each video. For each representation introduced, we also asked participants whether they found it readable, intuitive, and/or useful (on a Likert scale). We treat these questions as conversational triggers, and the answers as spontaneous responses. Contrarily, the fourth section asks participants to reflect on several important aspects of the DatAR system *after* they have seen them used in context during a mock-up user scenario.

This final section additionally contained sixteen bipolar adjective pairs, which aimed to measure the TAM constructs mentioned earlier as well as two antecedents to PU: Perceived Enjoyment (PE) and Perceived Informativeness (PI). We used the same adjectives as Rese et al. (2017) did in their analysis of augmented reality applications (see Figure 8). They found these adjectives to reflect the TAM model reasonably well, with the added benefit of indicating paths of improvement over more traditional item scales. To allow more granular analysis in our small sample, we deviated from Rese et al. (2017) by opting to ask users to fill in a bipolar scale instead of asking users to select a subset of adjectives; this is in line with the approach by Davis (1985).

Data Analysis

We performed an exploratory analysis on the quantitative data using SPSS (v26.0.0.0), and used Kendall's τ_b statistic to compare ordinal associations. Participants were instructed to give qualitative feedback throughout the survey; this was analysed using an inductive approach in Nvivo (v12.6.0.959). Responses were first open coded; codes were then aggregated where reasonable. Participants have been assigned numerical pseudonyms; to qualify responses, the nine participants who considered themselves neuroscientists (separate from their indicated discipline) received an additional

[21] Two participants did not fill in a discipline.

[22] More specifically: two current bachelor's students (who had also joined our second design cycle's evaluation), three participants with undergraduate degrees, five participants with graduate degrees, six current PhD students, one PhD, and five professors (of which three full professors).

[23] XR, or Cross Reality, encompasses both AR and VR.

Figure 6

Likert scale items posed during the video tutorials (section 3); primarily used as conversational triggers.

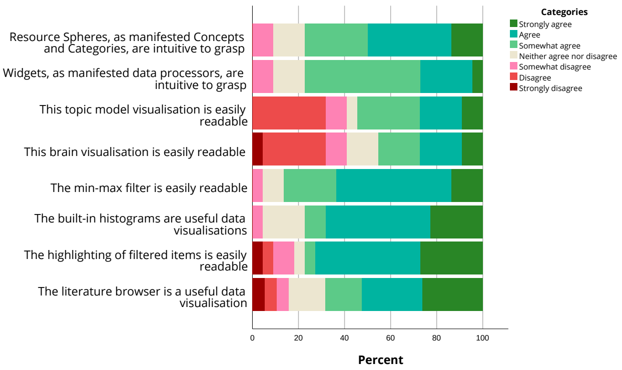
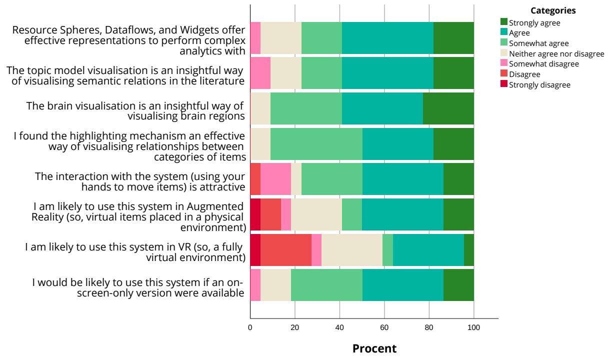


Figure 7

Likert scale items posed at the end of the survey (section 4); used to evaluate the system.



n designation (i.e., P1$_n$–P9$_n$ and P10-P22). Minor typos in responses were fixed for readability.

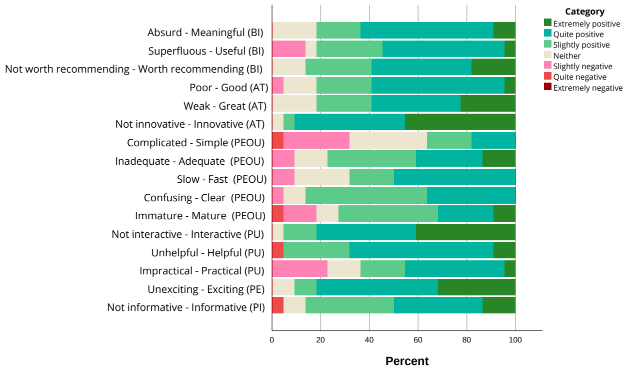
Results

Quantitative Results

We visualised participant responses on the Likert scale items shown during the video tutorials (Figure 6) and at the end of the survey (Figure 7), as well as on the bipolar adjective pairs (Figure 8). To qualify these aggregate visualisations, given the diversity of our sample, we checked whether subgroups significantly differed in their responses. We first investigated whether expertise in topic modelling impacted perception of our topic model visualisation; this was not the case ($\tau_b = -.063, p = .729$). Likewise, VR experience did not correlate with the likelihood of using DatAR in VR ($\tau_b = -.031, p = .862$), AR experience did not correlate with the likelihood of using DatAR in AR ($\tau_b = -.074, p = .734$), and neither VR nor AR experience correlated with likelihood of using DatAR on a 2D screen ($\tau_b = -.267, p = .066$; $\tau_b = -.115, p = .516$).

Figure 8

Bipolar adjective pairs (negative - positive); measures constructs of the TAM.



Note. Acronyms in parentheses refer to TAM components: Behavioural Intention to Use (BI), Attitude Toward Use (AT), Perceived Ease of Use (PEOU), Perceived Usefulness (PU), Perceived Enjoyment (PE) and Perceived Informativeness (PI).

We then looked at whether expertise in neurosciences or performing literature reviews caused different perceptions across all (24) non-demograhic response items. We found three significant correlations: participants knowledgeable in neurosciences were more likely to use the system in VR than those who were not ($\tau_b = .338, p = .006$); participants with a higher amount of experience in reviewing literature were more likely to (1) deem the system superfluous (rather than useful; $\tau_b = -.412, p = .002$) and (2) find using hand gestures to control the system attractive ($\tau_b = .323, p = .038$).

We will return to these findings and the quantitative responses more generally in the Discussion section. For an overview of all tests performed and their results, see Appendix F; given the high number of tests in this exploratory analysis, the p-values should be regarded critically.

Qualitative Results

Many participants took the time to write up additional feedback, qualifying their responses on TAM constructs and offering paths of improvement. In this section, we highlight recurring themes and critical insights.

Literature Review Practices. Thirteen participants (59%) mentioned using an academic search machine of some kind, including Google Scholar, PubMed, Web of Science, Scopus, PsycINFO, and DBLP. Four (18%) mentioned using boolean operators to improve their search queries. P8$_n$ used these "to capture publications that mentioned several concepts, so that the search is more specific."

The use of reference managers was also popular: Ten participants (45%) mentioned using software such as Mendeley, Nvivo, Endnote, and Excel to manage their readings. Five participants (23%) emphasised the importance of making notes and summarising papers, referring to tools such as

Microsoft Word, Evernote, and PaperShip.

Other practices mentioned were the use of review papers as starting point, snowballing (repeatedly following references in papers), using data sets, synthesising statistical findings across papers, and using concept-spotting tools (e.g., search-and-highlight in pdf files).

Use Cases. Throughout the survey, participants highlighted where it makes sense to use a system like DatAR – given more polish.

Five participants (23%) framed the system as useful during the early stages of a literature review process. P9$_n$ commented: "I would probably be interested in viewing data like this for relations I am relatively new to, just to get an idea, but if I am really doing a systematic review I would want to process them one-by-one and explicitly record the themes/dimensions." P1$_n$ took this one step further, placing the value of the system in use "during lectures and seminars in the beginning of the bachelors (...). In more specialized regions and higher education the students and teachers have a higher understanding of interspinal relations (...), wherefore this method could be innovative but also time consuming and therefore not as useful as for new students." P19 considered that if DatAR were applied to social sciences or humanities, it could be useful in avoiding tunnel vision by allowing exploration of terminologies across larger collections, though would require more manual work to extract meaning than in neurosciences (in which there is less contestation on terminology). P16 and P18 mused about the application of the system in educational sciences, with P16 stating they could "see a VR context with school intervention methods and its effect on child outcomes."

Others highlight areas in which the system could complement current practices (i.e., our *Augment, Don't Replace* design goal). P13 commented that the system has "great potential for complex information seeking tasks, although I would prefer traditional systems when my tasks are not as complex (e.g., lookup tasks)." P9$_n$ reflected more critically on where the system could fit in next to existing workflows (which we will return to in the Discussion section):

> My main concern is that I am having trouble imagining the type of review this would be worthwhile for. Most of the time we are researching something that we are already quite familiar with. We have already established an intuition of frontal vs. deep regions, and the visualization may not be so helpful. I feel this is a really helpful way to do a very broad and shallow review, but for me the most laboursome part of the process is retaining this sense of overview while processing the details. After I find a link between depression and the amygdala I would want to know what the papers did. I want to see what the imaging modality was, what the experimental design was, the number of subjects, get a sense of the quality of design/execution, how they analyzed the data, etc. And then after getting that, I would quite like to filter and zoom in and out, but for me, even though this looks exciting, excel would feel a lot easier and less laboursome.

Finally, P4$_n$ and P10 reflected on DatAR's lacking support for *systematic* literature review, wondering "what the next steps would be in order to write a review and organizing the literature based on these connections" and posing that "there needs to be a good way to quantify literature found."

Fundamental Concerns. While P4$_n$ felt XR visualisations had an advantage over the 2D screen, four other participants stated they were less convinced. P13 and P15 (both computer scientists) call for a comparison between our setup with an equally functional desktop set-up. P15: "Having the extra burden of VR/AR seems to me to be a distraction, somehow.. something that would tire me even more, allowing me to dedicate less time/attention to the actual information being analyzed." These participants also pointed out some more fundamental issues with using a three-dimensional topic model. P13 observes that items at its centre are inherently harder to access than those at the periphery; P15 lists several cognitive biases when reading 3D clouds, which need to be mitigated by the system to guide correct understanding, and wonders whether an abstract document space would be worth this trouble; P4$_n$ worried about decreased legibility if more concepts had to be taken into account in the topic model than was the case.

Suggestions for Improvement. Other than concerns about the system's core design decisions, we also received many suggestions for improvements. We will cover suggestions recurring at least twice here; other comments have been collected in Appendix G.

P9$_n$, P11, and P14 found the colour scheme insufficiently intuitive, proposing alternative scales. P15 and P16 argue for using 2D UI elements where possible, such as with the *Min-Max Slider*. This would allow having the UI parallel with the view plane, improving the readability of text.

A few participants worried about the information density of the two visualisation Widgets, calling them "overwhelming" (P12), "easy to get lost [in]" (P14), and "a little vague as there was so much in there" (P4$_n$). P2$_n$ elaborated: "I find it very understandable, but not actually readable because all the regions are so close together."

Another recurring theme concerns the valence and qualitative relationship of co-occurrences (P4$_n$, P8$_n$, P12, & P19). The current system does not distinguish between positive and negative relationships, nor other types of association. To move beyond "a useful screening tool in an early phase" (P12), this will need to be supported.

Methodological Comments. Participants commented on the video survey too: Some found the voice-over hard to understand, one found it necessary to pause the video to catch some visuals, and one reported nausea due to the VR recording. There were also some comments on phrasing and a suggestion to integrate the tutorials in the scenario. More universal were comments either on shortcomings of the (non-interactive) XR video evaluation format, or misconceptions caused by it. For example, it being unclear how to select spheres (by grabbing), a complaint about the removal of a RS from the Brain Region Visualisation (which was actually duplicated rather than removed), and a complaint about using a rendered laptop screen in VR instead of a viewpoint-aligned display (which, as explained to participants, was there to simulate an AR environment – which is harder yet to get across on video).

Discussion

In this evaluation study, we set out to investigate whether researchers would see themselves adopt DatAR as a serious tool in their literature review practices. On the positive side, our results indicate that participants thought the DatAR system was innovative, interactive and exciting. A large majority found the representational design easy to grasp and an effective means of performing complex analytics. On the negative side, participants were concerned about DatAR's complicated design, and (less so) about its impractical nature and immature state. Looking at these findings through the lens of the Technology Acceptance Model, it appears that the most major concern is Perceived Ease of Use. This finding is mirrored by conflicting responses on the Topic Model and Brain Region Visualisations: readability of these were reviewed negatively by almost half of the participants, while general usefulness of the approaches behind these two visualisations (later asked at the end of the survey) were not once reviewed negatively. Together, these findings suggest the merit of our general approach to the analysis of document collections, as well as our representational design choices, but raise serious questions about our visualisation design.

These quantitative findings are echoed in the qualitative feedback, where two camps can be discerned: those who see aspects of the system that can be improved, and those who see more fundamental flaws inherent to the AR platform. To start with the former: There is certainly room for improvement in fine-tuning the visualisations, considering such aspects as improved colour schemes and rethinking how to deal with an overwhelming number of data points. This latter aspect was somewhat intentional, given our "highlight, not hide" design requirement. The greatest trouble we faced was getting across concept names in the Topic Model Visualisation. Unlike the Brain Visualisation (fitting in SciVis), a topic model has no inherent spatial organisation (fitting in InfoVis). Our current solution – showing user-facing labels on all Resource Spheres – was generally perceived as overwhelming, and is inherently unscalable when the amount of data would increase even further. Implementing a metaphorical fisheye lens might be considered, emphasising that – while all data is visible – some data is at the centre of attention. This is especially difficult in XR, with its infinite spatial canvas, but some work has been performed on graphs in this area (see Kwon et al., 2016). It would be worthwhile to perform comparative research into alternative solutions, including the use of other visual parameters than opacity and colour.

Platform-level critiques of DatAR raise a more fundamental question: Are the solutions offered by Immersive Analytics indeed a good fit for the activity we are trying to support? We cannot ignore that more participants rated themselves likely to use the system if it were fully available in a desktop environment than if it would be offered in AR (or VR). Likewise, we found that scholars experienced in literature review were more likely to find our system superfluous (though, to a lesser degree, also deemed hand gesture control more attractive than others). While there are other factors at play here (e.g., the AR platform was novel to most users and using a video survey format has some inherent drawbacks in getting across the XR experience), these findings do raise existential questions for our project that need to be addressed. One direction in which we are trying to address this in our research group is by exploring an alternative to free-floating AR: Augmented Displays (Reipschläger & Dachselt, 2019). In particular, we are looking into extending the mouse cursor into augmented reality space. This has its own challenges, but may alleviate the issue of switching between two control modalities. To rhyme this with our design requirement making use of the medium," we consider Augmented Displays (as a hybrid AR environment) complementary to the full AR version of the system. These can live side-by-side, depending on where the users' attention is required – fully on their desktop screen, in AR, or perhaps in the environment at large. That said, we concur that comparative research will be required to make any conclusive statements on whether AR is an appropriate platform for relation-finding in document collections, and will be working towards this.

Functionality-wise, we must concede that there is currently insufficient augmentation of current work practices (the design requirement "augment, not replace"). Participants noted the system's usefulness in the early stages of literature discovery (which is what it was designed for), but criticise how findings cannot find their way to their (manual or computer-supported) reference management systems. In other words, just sending data over to the desktop environment is not sufficient – integration with existing software would be the next necessary step to augment current practices. Reflecting on the comments by P9_n, it would be particularly fruitful to consider the reverse direction of what we currently implemented: to take in an already existing overview of papers and allow researchers to visualise and expand the scope of this overview using DatAR. This two-way street between the desktop and AR platforms requires more attention.

Other than comments on the current system, participants also offered ways to extend DatAR beyond its original purpose. Particularly interesting was the suggestion to use the system in educational settings, as a means to get across the complex web of relationships in the literature to bachelor's students. While we focused our design on supporting subject experts so far, we did muse about the possibility of educational use of our system. It would be worthwhile to explore how the system could be tailored to this, for example, by facilitating *guided* explorations of the data in alignment with the curriculum.

An advantage of our diverse participant pool was hearing their thoughts on how the system could transfer to other disciplines than neuroscience. While the brain visualisation was developed specifically to support neurocientists, all other Widgets, data structures and algorithms are domain-generic: Given equally-structured co-occurrence data and a topic model, the system would be able to visualise another document collection. However, this technical feasibility does not directly translate to a *useful* system for practitioners. Once DatAR has been further improved within the neuroscientific domain, it would be highly valuable to study whether it could offer value in other domains – in particular, whether the epistemic design still holds for other disciplines' task requirements.

Reflections on Rigour

After having written this discussion, we reflected on each of Meyer and Dykes' (2019) criteria for rigour. This reflection can be found in Appendix H. We hope this allows readers to place this work in its research context.

Conclusion

Maintaining an overview of publications in the neuroscientific field is challenging, especially with an eye to finding relations – both commonplace and novel. To support neuroscientists in this challenge, we developed a novel analytics system, DatAR, that combines work in Text and Immersive Analytics to find ways to make automated methods more accessible and integrated with current literature review practices. After formulating and implementing the design requirements, we evaluated this system using a video survey.

In the results, we see a discrepancy between participants' perception of the general analytics approach and the use of the XR platform to embody this approach. This may partly be explained by the use of recordings of a mocked AR scenario rather than having participants experience the system first-hand. However, participants also pointed out more fundamental issues with our chosen 3D visualisations, and the currently problematic integration of shallow relation-finding (in AR) and deep sensemaking (on desktop) – both issues that warrant addressing. Conversely, the AR-tailored epistemic and representation design of DatAR were generally perceived as suitable for performing complex analytics. We see opportunities in pushing the Resource Sphere–Widget–Dataflow paradigm to its limits by expanding their scope in augmenting intellectual activities.

All-in-all, the reported design cycle and its evaluation study have generated sufficient fuel for a continued exploration of the fundamental question in our work: Is AR a suitable medium for relation-finding in document collections, and – if so – how? The natural trajectory of this research agenda is towards a comparative study, looking at AR in comparison to the desktop environment given the same task. However, before a *fair* comparison can be made, addressing the perceived issues with ease of use in DatAR will likely not be sufficient: AR technology itself needs to have matured further too. Until then, the DatAR system has a useful role to play in speculating on the future of human-computer interaction and the willingness of practitioners to adopt AR technologies in their work.

References

Antoniou, G., Groth, P., van Harmelen, F., & Hoekstra, R. (2012). *A semantic web primer* (3rd). MIT Press.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., & Vidal, M. E. (2018). Towards a knowledge graph for science, In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, New York, NY, USA, ACM. https://doi.org/10.1145/3227609.3227689

Bakker, A. (2018). *Design research in education: A practical guide for early career researchers*. London, UK, Routledge.

Bron, M. (2013). *Exploration and contextualization through interaction and concepts* (PhD Thesis) [https://hdl.handle.net/11245/1.399870]. University of Amsterdam. https://hdl.handle.net/11245/1.399870.

Cater-Steel, A., Toleman, M., & Rajaeian, M. M. (2019). Design science research in doctoral projects: An analysis of australian theses. *Journal of the Association for Information Systems*, *20*(12), 1844–1869. https://doi.org/10.17705/1jais.00587

Chuttur, M. (2009). Overview of the Technology Acceptance Model: Origins, developments and future directions [https://aisel.aisnet.org/sprouts_all/290]. *Sprouts: Working Papers on Information Systems*, *9*(37).

Cordeil, M., Dwyer, T., Klein, K., Laha, B., Marriott, K., & Thomas, B. H. (2017). Immersive collaborative analysis of network connectivity: CAVE-style or head-mounted display? *IEEE Transactions on Visu-*

alization and Computer Graphics, *23*(1), 441–450. https://doi.org/10.1109/TVCG.2016.2599107

Cordeil, M., Cunningham, A., Dwyer, T., Thomas, B. H., & Marriott, K. (2017). ImAxes: Immersive axes as embodied affordances for interactive multivariate data visualisation, In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, ACM. https://doi.org/10.1145/3126594.3126613

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* (PhD Thesis). Massachusetts Institute of Technology.

Ens, B., Anderson, F., Grossman, T., Annett, M., Irani, P., & Fitzmaurice, G. (2017). Ivy: Exploring spatially situated visual programming for authoring and understanding intelligent environments, In *Proceedings of the 43rd graphics interface conference*, Waterloo, CAN, Canadian Human-Computer Communications Society.

Federico, P., Heimerl, F., Koch, S., & Miksch, S. (2017). A survey on visual approaches for analyzing scientific literature and patents. *IEEE Transactions on Visualization and Computer Graphics*, *23*(9), 2179–2198. https://doi.org/10.1109/TVCG.2016.2610422

Fonnet, A., & Prié, Y. (2019). Survey of immersive analytics. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2019.2929033

Gopalakrishnan, V., Jha, K., Jin, W., & Zhang, A. (2019). A survey on literature based discovery approaches in biomedical domain. *Journal of Biomedical Informatics*, *93*, 1–18. https://doi.org/10.1016/j.jbi.2019.103141

Görg, C., Tipney, H., Verspoor, K., Baumgartner, W. A., Cohen, K. B., Stasko, J., & Hunter, L. E. (2010). Visualization and language processing for supporting analysis across the biomedical literature (R. Setchi, I. Jordanov, R. J. Howlett, & L. C. Jain, Eds.). In R. Setchi, I. Jordanov, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-Based and Intelligent Information and Engineering Systems*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-642-15384-6_45

Gracia, A., González, S., Robles, V., Menasalvas, E., & von Landesberger, T. (2016). New insights into the suitability of the third dimension for visualizing multivariate/multidimensional data: A study based on loss of quality quantification. *Information Visualization*, *15*(1), 3–30. https://doi.org/10.1177/1473871614556393

Gramatica, R., Di Matteo, T., Giorgetti, S., Barbiani, M., Bevec, D., & Aste, T. (2014). Graph theory enables drug repurposing – how a mathematical model can drive the discovery of hidden mechanisms of action (R. Lambiotte, Ed.). *PLoS ONE*, *9*(1), 1–10. https://doi.org/10.1371/journal.pone.0084912

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–356. https://doi.org/10.25300/MISQ/2013/37.2.01

Hevner, A. R. (2007). A three cycle view of design science research, *19*(2), 1–6.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research [https://www.jstor.org/stable/25148625]. *MIS quarterly*, *28*(1), 75–105.

Kerzner, E., Goodwin, S., Dykes, J., Jones, S., & Meyer, M. (2019). A framework for creative visualization-opportunities workshops. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 748–758. https://doi.org/10.1109/TVCG.2018.2865241

Kwon, O.-H., Muelder, C., Lee, K., & Ma, K.-L. (2016). A study of layout, rendering, and interaction methods for immersive graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, *22*(7), 1802–1815. https://doi.org/10.1109/TVCG.2016.2520921

Lander, A. D. (2010). The edges of understanding. *BMC Biology*, *8*(1). https://doi.org/10.1186/1741-7007-8-40

Li, Z., Zhang, C., Jia, S., & Zhang, J. (2019). Galex: Exploring the evolution and intersection of disciplines. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1182–1192. https://doi.org/10.1109/TVCG.2019.2934667

Liu, H., Liu, C., & Belkin, N. J. (2019). Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology*, *56*(1), 166–175. https://doi.org/10.1002/pra2.63

Mäkelä, V., Radiah, R., Alsherif, S., Khamis, M., Xiao, C., Borchert, L., Schmidt, A., & Alt, F. (2020). Virtual field studies: Conducting studies on public displays in virtual reality, In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, Association for Computing Machinery. https://doi.org/10.1145/3313831.3376796

Mankoff, J., Rode, J. A., & Faste, H. (2013). Looking past yesterday's tomorrow: Using futures studies methods to extend the research horizon, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. https://doi.org/10.1145/2470654.2466216

Marriott, K., Chen, J., Hlawatsch, M., Itoh, T., Nacenta, M. A., Reina, G., & Sturzlinger, W. (2018). Immersive Analytics: Time to reconsider the value of 3d for information visualisation (K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, & B. H. Thomas, Eds.). In K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, & B. H. Thomas (Eds.), *Immersive Analytics*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-01388-2_2

Marriott, K., Schreiber, F., Dwyer, T., Klein, K., Riche, N. H., Itoh, T., Stuerzlinger, W., & Thomas, B. H. (Eds.). (2018). *Immersive analytics* [https://www.springer.com/gp/book/9783030013875]. Springer International Publishing.

Meyer, M., & Dykes, J. (2019). Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 87–97. https://doi.org/10.1109/TVCG.2019.2934539

Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921–928. https://doi.org/10.1109/TVCG.2009.111

Nunamaker, J. F., Jr., & Briggs, R. O. (2011). Toward a broader vision for Information Systems. *ACM Transactions on Management Information Systems*, *2*(4), 1–12. https://doi.org/10.1145/2070710.2070711

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, NY, Basic Books.

Reipschläger, P., & Dachselt, R. (2019). DesignAR: Immersive 3D-modeling combining augmented reality with interactive displays, In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces - ISS '19*, Daejeon, Republic of Korea, ACM Press. https://doi.org/10.1145/3343055.3359718

Rese, A., Baier, D., Geyer-Schulz, A., & Schreiber, S. (2017). How augmented reality apps are accepted by consumers: A comparative analysis using scales and opinions. *Technological Forecasting and Social Change*, *124*, 306–319. https://doi.org/10.1016/j.techfore.2016.10.010

Rizzo, A., & Bacigalupo, M. (2004). Scenarios: Heuristics for action. *Proceedings of XII European Conference on Cognitive Ergonomics*, 153–160.

Rogers, Y. (2012). HCI theory: Classical, modern, and contemporary. *Synthesis Lectures on Human-Centered Informatics*, *5*(2), 1–129. https://doi.org/10.2200/S00418ED1V01Y201205HCI014

Sedig, K., Parsons, P., & Babanski, A. (2012). Towards a characterization of interactivity in visual analytics. *Journal of Multimedia Processing and Technologies*, *3*(1), 12–28.

Sicat, R., Li, J., Choi, J., Cordeil, M., Jeong, W., Bach, B., & Pfister, H. (2019). DXR: A toolkit for building immersive data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 715–725. https://doi.org/10.1109/TVCG.2018.2865152

Stravinsky, I. (2003). *Poetics of music in the form of six lessons*. Harvard University Press.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18. https://doi.org/10.1353/pbm.1986.0087

Tanhaei, G., Hardman, L., & Huerst, W. (2019). DatAR: Your brain, your data, on your desk - a research proposal, In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. https://doi.org/10.1109/AIVR46125.2019.00029

Thomas, B. H., Welch, G. F., Dragicevic, P., Elmqvist, N., Irani, P., Jansen, Y., Schmalstieg, D., Tabard, A., El-Sayed, N. A. M., Smith, R. T., & Willett, W. (2018). Situated Analytics (K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, & B. H. Thomas, Eds.). In K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, & B. H. Thomas (Eds.), *Immersive Analytics*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-01388-2_7

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

W3C. (2013). SPARQL 1.1 query language.

Woods, D. D., Patterson, E. S., & Roth, E. M. (2002). Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition, Technology & Work*, *4*(1), 22–36. https://doi.org/10.1007/s101110200002

Appendix A
Introduction to Appendices

In these annotated appendices I offer an overview of work performed that did not make it (fully) into this thesis, including documentation, source code, analysis results, data mappings and designerly reflections. Some of these documents are publicly available and therefore linked. Others are stored in the archives of the DatAR research group; access can be requested from Ghazaleh Tanhaei (Utrecht University).

Appendix B
Source code & Contributor's Guide

Over the course of the project, we've developed two main programs: DatAR Unity, and DatAR Web. These have been carefully version-controlled using Git and adhere to consistent coding styles. Moreover, licenses have been taken into account to allow eventual open-sourcing of the code base. Next to these, I've written a Jupyter notebook to generate topic models based on data in the BAG. We did not use this data in the end, instead opting to use the data from Cunqing Huangfu (BAG custodian). In the archive these files and external references can be found under the header "Source Code: Overview June 2020."

Next to this, an extensive contributor's guide has been written. This guide is available here and contains more details on the buildings blocks of the DatAR system.

Appendix C
Data Mappings & Infrastructure

In order to link the different data sources (BAG, Wikidata, MeSH, Scalable Brain), additional mappings were performed by Tessa Heeroma and Anna van Harmelen. I converted these mappings to RDF and uploaded these to a private triple store (running Virtuoso 7.2.5). All coordinate data (for brain regions and the topic model) were also uploaded there. To request access to this data store or a data dump of these mappings (archived under "Data mapping: Overview June 2020"), you may contact us.

Appendix D
Design & Research Documentation

Our process was recorded in two main ways: in a document database and in a wiki. The document database contains 77 entries, containing minutes of supervisory and design meetings, important correspondences within the team and with partners, documentation of evaluation studies, video and image exhibits, and proposal/planning documents. All these have been tagged with dates of creation and participants. All documents mentioned in this section have also been stored here. Whereas these documents are static snapshots, the wiki (structured as a Zettelkasten) was meant to be dynamic. It contains 30 entries, ranging from indices, summaries on themes in the literature (e.g., text readability in AR), a coding style guide, a repository of SPARQL queries, and paper digests. Moreover, I kept a weekly research logbook on activities I undertook. All this documentation was performed to ensure a traceable design process, allowing the depth of reporting envisioned by Meyer and Dykes (2019).

Appendix E
End-of-Project Reflection

Before writing up the Discussion & Conclusion sections of this paper, I went through all of the archive and wrote up an extensive reflection document (approx. 2750 words). This document contains four sections. "Origins & Rationales" traces several design decisions over the time of the project; "Open Challenges" outlines several recurring issues that will need to be addressed in the follow-up to my contribution; "Further Research and Loose Ends" summarises all directions we pursued but had to drop at some point; "Process Reflection" then offers a (self-)critical reflection on the design process as a whole. Little of this document ended up in this thesis, given most points made are relevant internally rather than for the scientific community as a whole. This document is archived under "Reflections on Design Process: Overview June 2020." We offer an abridged version of the Process Reflection here:

> Given the many factors involved in building a novel system to study willingness to adopt AR technology to support relation-finding, the DatAR project has often needed to reinvent itself. This search for meaning is somewhat inherent to technology-oriented questions, and has been the factor driving this project. At the same time we recognised that the success of our work depended primarily on one thing: whether the task, relation-finding, was effectively supported by our environment. There is a tension between these two concerns – working from task requirements and from technological possibilities. There is an upside to such artificial boundaries in problem-solving, however. Composer Igor Stravinsky once noted on his creative process that "my freedom will be so much the greater and more meaningful (...) the more I surround myself with obstacles" (2003, p. 65). By creating an interface with essentially one interaction – to grab – we were forced to break down the tasks to their bare essence on the interaction level. I wonder whether these same representations could be transferred to other interface paradigms in the future too...

Appendix F
Video Survey: Scenario, Questions, Videos, Results

To perform our evaluation study, we developed a tutorial series to introduce participants to our system as well as a more free-form user scenario based on an authentic task. We wrote a script for the tutorial series, voiced the lines, and subsequently recorded the steps inside of the DatAR environment. The user scenario has a defined start, but the rest was improvised. The survey was built using the Qualitrics survey platform. We embedded the videos directly in the survey. A document with all questions, as well as the script and the videos, is accessible here. For the sake of transparency, we also attached results from all statistical tests there.

Appendix G
Video Survey: Additional Suggestions for Improvement

Other than reported in the paper, individual participants have noted several additional ways to improve the DatAR system. I list these here, including some discussion where useful:

- Several participants offered ways in which Resource Spheres and Widgets could be extended in functionality. P14 noted that finding individual RSs in visualisation Widgets (e.g., to create a duplicate) would be challenging. This suggests the need for a local or global RS highlighting mechanism (based either on the currently held RS, or another RS selection mechanism). P13 notes that placing RSs in Inspector Widgets would quickly become tiresome, and wonders whether a hotkey could be added; likewise, they note that being able to easily duplicate dataflows would be useful. P18 suggests a more extensive change: "Similar to a mind-map, I expected the Resource Spheres to be larger and branching off into different things." Given these responses, it would certainly be worth considering how RSs could be made more interactive on their own; on the other hand, we should be careful to not complicate the task division between RSs and Widgets too much.
- P11 (a statistician) wonders to what extent the tool could be customised. As we built the system modularly, it should be possible for data scientists to write their own extensions and (partially) move their analytics workflow to AR. While we focused on supporting domain experts, it would be interesting to see what data analysts could do with this system given access.
- P11 muses that combining two visualisations ("two brain graphs") could be useful. During our design process, we did consider the idea that visualisation Widgets could be piped into meta-visualisations on which users could perform set operations (i.e., union, difference, intersect). This idea is still under consideration for a next study.
- P13 would like to see "color-coded clustering for disease types." There is currently no data in the system that categorises diseases into subgroups, though this could be a worthwhile expansion in the future. More generally, disease–disease comparisons could be a next area of support.
- The current system is built around co-occurrences of two concepts. P13 wonders if it would be possible to look for three (or more) concepts co-occurring at once. It would be interesting to experimentally test this idea in a next design phase, see what happens, and – most importantly – come to a better understanding of what insights can be gained with such a set-up.
- P12 criticised the display of source materials in the DatAR Web Companion, suggesting that the system should show "3-5 sentences before and after the sentence in question, to optimally show the context of the claim."
- On the interaction-side, P8$_n$ saw that exact selection of parameters in the Min-Max Filter Widget was challenging, and wondered if it would be "possible to zoom in/select more carefully?"

Finally, we received an additional direction for future research: What the effect of this system would be on RSI (P11).

Appendix H
Reflections on Rigour

The following aspects are the criteria for rigour set out by Meyer and Dykes (2019). We shortly reflect on each of these.

Informed. In the early stages of our design process, we sought clarity on our epistemological position – deciding on a constructivist/constructionist stance (as discussed in detail in the Methods section). We also started up each design cycle with a new round of literature review, keeping up-to-date with the latest developments and digging into new areas of relevance.

Reflexive. Next to keeping a logbook, at the end of the project I wrote a final reflection on the project from a personal standpoint (see Appendix E for an abridged version). We also framed our evaluations as a conversation, opening ourselves up to critical dialogue. Finally, the research group presented a work-in-progress version at the AIVR 2019 conference (Tanhaei et al., 2019).

Abundant. While our contact with participants in the current evaluation study was minimal, we tried to faithfully portray their voices by using their own words where possible. We also made the decision to invite a diverse group of participants to comment on our work, which yielded interesting new perspectives to consider. However, the sample size of 22 cannot yet be considered sufficient to draw generalisable conclusions.

Plausible. We have been careful to phrase our conclusions in terms of our own system design and its particular niche: supporting relation-finding in neuroscientific text collections using Augmented Reality. We also zoomed out to discern possible mechanics that might work outside the scope of our particular niche, drawing on the building blocks of our representational design. On request, we are also happy to share archival materials of our design process.

Resonant. While resonance is more in the eye of the beholder (and therefore hard to judge at this stage), Meyer and Dykes (2019) do note that *transferability* could be considered a form of resonance. We hope the multidisciplinary voices of our participants brought in ideas for colleagues of various neighbouring fields.

Transparent. We have tried to offer a detailed account of our results, as well as a rich set of appendices. The contributor's guide (Appendix B) and materials related to the video survey (including source files; Appendix F) have been made publicly accessible. Furthermore, we have tried to convey our process using this reflection on rigour, as well as offering an abridged version of the general process reflection (Appendix E). Other access can be arranged on request.

Appendix I
Provenance and Embodiment of design requirement.

Table I1 contains an overview of the design requirements, their origins, and how they have been implemented in the final system. For context, we also added a visual snapshot of the interface at various stages in the design process (Figures I1– I4).

Figure I1

Version 2019.0.0, used during the evaluation of the first design cycle.

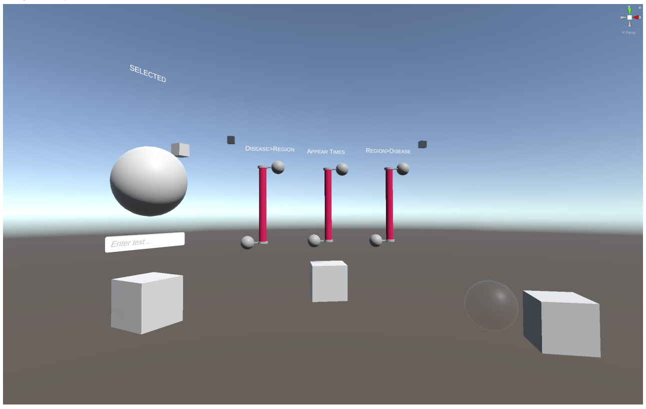


Figure I2

Version 2020.0.0, used during most evaluations of the second design cycle. This is the first version that connected with the DatAR web companion.

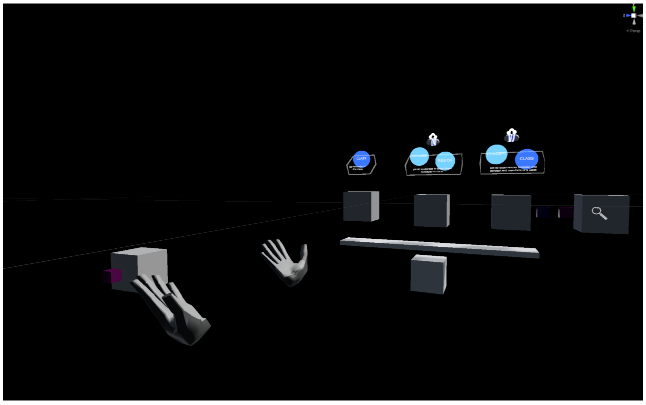


Figure I3

Version 2020.0.1, used during the final evaluation round of the second design cycle – included several fixes after initial feedback.

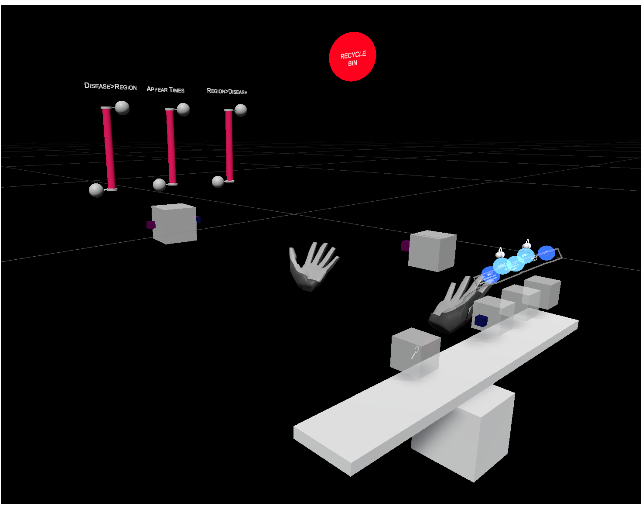


Figure I4

Version 2020.1.0, used during the evaluation of the third design cycle, presented in this thesis. Note that a VR virtual environment was used to mock the AR situation.

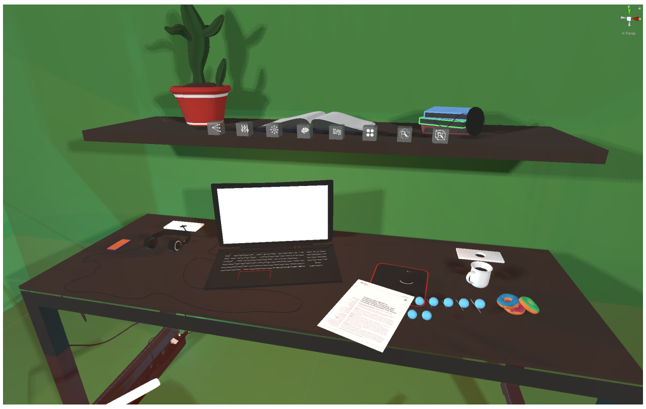


Table I1

Provenance and Embodiment of Design Requirement. We refer to the three design cycles using C1, C2 and C3.

Design requirement	Provenance	Embodiment
Supporting open and closed discovery	We identified this design requirement during the initial literature review of C1. Based on user feedback during C2, we added Inspect as a necessary component in our epistemic design.	In our epistemic design, we embodied this requirement in the Contextualise task: We utilised domain semantics to allow users to initiate discovery of meaningful direct relationships (co-occurrences). More work is required to extend the system to support indirect relation-finding and fulfil the promise of this design requirement fully.
Highlight, not hide	This requirement had been implicit in our work ever since we implemented the Min-Max Filter and its highlighting behaviour (C1), but only became explicit during a later literature review (C3). This was after one of our participants called our non-optimised topic model visualisation a "trashy Christmas tree" (C2:P7).	The filter behaviour of the Topic Model and Brain Region Visualisations (which highlight rather than hide) were based on this requirement. Until C2, we used colour coding to distinguish filter states. After C2, we increased the contrast between highlighted and non-highlighted RSs by using opacity (without making any RSs illegible). We made highlighting behaviour system-wide when we added the Brain Region Visualisation during C3.
Augment, not replace	The more we discussed current literature review workflows with practitioners (C1 and C2) and reflected on our own practices, the more it became clear that our system would not be able to replace these without unacceptable compromises. Given Augmented Reality allowed us to, we decided to design for a blended approach: extending and supporting existing workflows by our system's functionality.	After processing the results of C1, we decided to implement the DatAR Web Companion as a bridge between current screen-based workflows and the relation-finding capabilities of the DatAR system. In our third design cycle, we additionally alluded to a situated analytics feature: extracting concepts from a printed academic paper in the users' view. Given the results of C3, work is left in integrating with existing digital tools more tightly (such as reference managers and search machines).
Making use of the medium	During C1, we developed a Search Machine Widget that required the keyboard. We found that keyboard focus in 3D space was tricky. Both using gaze and hand tracking to set focus had disadvantages. Moreover, after we had implemented the web companion to satisfy the above requirement, we found it unergonomic to share a keyboard between both the AR interface and the regular desktop environment. We removed the keyboard from the system during C2, during which we implemented new query widgets that worked on categories rather than plain text searches. The other aspect of this design requirement, to build a decentralised interface, was discovered while building the prototype during C1. Originally, the system's Widgets were already connected prior to any user interaction; playing around with the (then invisible) dataflows allowed more complex constellations of filters and visualisations, which we found promising.	In our current system, all interactions take place using a simple grabbing gesture. The keyboard is dedicated to controlling the regular desktop environment, optionally including the DatAR web companion. All Widgets process their data locally, can be placed anywhere, and can be connected freely – realising the vision of a decentralised interface. An added benefit we found is that this architecture made it easy to build new Widgets by combining atomic building blocks (e.g., Inlets, Outlets, Receptacles, and RS Manufacturers).

Appendix J
Design Research: An Overview

The following is a digest on design research. It elaborates on the frameworks that shaped our our design practices. Design *research* distinguishes itself from routine design by contributing to an (academic) knowledge base. Routine design generally applies known solutions to known problems, whereas design researchers choose areas in which the problem and/or solution are novel (Gregor & Hevner, 2013). Design research can be viewed as a methodological framework, "a set of approaches with family resemblances" (Bakker, 2018), which means that there are diverse opinions on what counts as a good design study. Indeed, design research can be approached from multiple epistemological and ontological mindsets, as well as both quantitatively and qualitatively (Meyer and Dykes, 2019; Hevner et al., 2004).

DatAR exists in the overlap of multiple disciplines, each of which developed an approach to design research. It was therefore important to critically assess which design research paradigm we would place our work in. In this appendix, we cover two perspectives on design research: from Information Systems and Information Visualisation.

Insights from Information Systems

Information Systems (IS) research is focused on the use of information technology to aid organisational and business purposes (Hevner et al., 2004). While DatAR does not fit under that label (although the field has been moving towards broadening its scope to include, e.g., bioinformatics, cf. Nunamaker and Briggs, 2011), the seminal work by Hevner et al. (2004) still offers helpful guidelines in understanding and applying design research methods. These authors argue that knowledge-gathering in IS requires two "complementary but distinct paradigms" (p. 3): behavioural science – which attempts to explain or predict phenomena and aims to generate *understanding* – and design science – which attempts to change phenomena and aims to generate *utility*. Informed by the environment in which the information system is placed and the existing knowledge base, design artefacts can be built and evaluated in order to offer utility to the environment as well as add to the knowledge base. This synergy between relevance (in the environment), rigour (in touch with the knowledge base) and the design process itself is crucial for high-quality design research (Hevner, 2007). An artefact (or, design product) is a broad term that can refer to (1) a *construct* that provides vocabulary to discuss matters, (2) a *model* that allows for representing a problem domain, (3) a *method* that prescribes how to perform something, and (4) an *instantiation* which is an implemented (prototype) system. According to the authors, each of these artefact types should be recognised as valid design contributions.

Gregor and Hevner (2013) furthermore argue that while a well-developed design theory is the optimal type of contribution to arise from a design study, it is unrealistic to expect researchers to get to that stage without first developing situated implementations of artefacts and formulate *nascent* design theories. As a particular research area matures, more descriptive and prescriptive knowledge is added to the knowledge base. This raises expectations as to what counts as a contribution. When both solution maturity and application domain maturity are high, we are talking about *routine design* (which is not promising for research). When we develop new solutions for known problems, we engage in *improvement*. Reversely, when known solutions are applied to a new problem space, we engage in *exaptation* (a term borrowed from biological evolution that describes a trait being adapted for a different purpose than its original one). In the rare case both the problem and the solution are novel, we can call this genuine *invention*. Hevner et al. (2004) and Gregor and Hevner (2013) also offer practice-oriented guidelines on how to frame, perform and write up design research.

Cater-Steel et al. (2019) analysed 40 Australian doctoral theses that engaged in IS design research. One troubling finding was that nearly half of these theses failed to discuss their research philosophy in terms of ontology or epistemology, "suggesting a lack of sophistication in the research process of these students" (p. 1854). More generally, the language on what design research is exactly (e.g., a paradigm or methodology) was muddled – even though scholars in the field have not reached a shared consensus on this either, it would have been better to explicitly pick one of the schools of thought. 60% of theses did not develop new theory nor extended or reexamined current theory. As mentioned already, Gregor and Hevner (2013) have argued that this is not a necessity for a useful contribution; Cater-Steel et al. (2019) muse that this lack may be due to time constraints of PhD candidates.

Insights from Information Visualisation

A field even more closely aligned with DatAR is that of Information Visualisation (InfoVis). Design studies arguably form the backbone of the InfoVis field, which has yielded many method-focused papers to draw from. Munzner (2009) offers a nested model to consider design and validation of a particular visualisation strategy. At the top layer we find the domain problem characterisation. This concerns the high-level domain-specific tasks users perform in the target domain and the data they use. Without a clear picture of these two aspects, a designer runs the risk of solving the wrong problem. The second layer is the operation and data type abstraction. Operations are understood as low-level domain-generic tasks (such as comparing and querying); the data type concerns how to represent the raw data (such as in a table or graph). A mismatch in this area runs the risk of showing users the wrong thing for their purpose. The third layer includes visual encoding and interaction design.

Without doing this right, the effective communication of the data is at risk. Finally, the fourth layer is algorithmic: how well the implementation is optimised. The authors argue that the nested layers are interdependent; e.g., a perfect visualisation is useless if it fails to solve the problem of interest. Whereas these authors primarily look at the process of design, Meyer and Dykes (2019) looked at how to safeguard rigour. Meyer and Dykes (2019) position themselves as interpretivist and look at criteria a work can be judged by after the design process has completed and has been written up. The authors expect a good design study to be: (1) informed by already existing knowledge, (2) reflexive of the researchers' own role in the study, (3) abundant in having considered and tried many possibilities, and using thick description to convey information, (4) plausible in making knowledge claims that are evidence-based, context-aware and persuasive, (5) resonant by being transferable and evocative, and (6) transparent in being particular enough about reporting.