V. Future data

How does one describe data and the work they do? How does one get access to data and strategies for their characterization? In an interview with simulation modeler Jana Solberg, she tells me about her work as a member of the modeling team generating the SSPs. The purpose of the latter is to serve as a standardized set of socioeconomic storylines that can 'drive' all global climate-modeling endeavors in the world. When Jana was talking about the SSPs, she pointed at elements displayed on the website, data repository and viewer *SSP database* – file and folder structures, diagrams and textual descriptions. After the interview, I come back to the database to learn more about the SSPs and their representation as digital datasets.

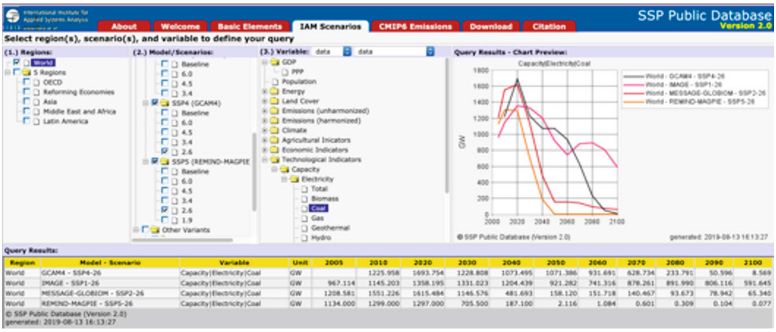
Access log

Based on the information obtained during my interview with Jana, I type 'ssp database' into Google's query tool[72] and click on the first of the search results, forwarding me to a bulky URL.[73] The main dashboard is hidden by a pop-up window displaying an agreement to Terms of Use. It includes terms on copyrights, citations required and liabilities. The agreement aims at protecting the provider of the online repository (here IIASA, the International Institute for

[72] Or rather www.qwant.com or www.startpage.com, retrieved on April 3, 2019.

[73] https://secure.iiasa.ac.at/web-apps/ene/SspDb/dsd?Action=htmlpage&page=about, retrieved on April 3, 2019.

Applied Systems Analysis) and the developers of the datasets (e.g. the PIK). Clicking 'I agree to the Terms of Use,' a next checkpoint is waiting for me: "Please use this button to log in as a guest user (restricted preview) or use the form below (and your individual email and password) to log in with your email." Doing the latter, I finally receive access to the website. The design of the website is very simple, with links to subpages such as 'welcome,' 'basic elements,' 'IAM scenarios,' 'CMIP6 Emissions', 'download' and 'citation.' I browse through the content and get caught by 'IAM scenarios,' identifying it as the cornerstone of the website. The subpage is structured as a so-called data viewer, representing the datasets of the SSPs in various forms (see Figure 37).



Select region(s), scenario(s), and variable to define your query

Query Results:

Region	Model - Scenario	Variable	Unit	2005	2010	2020	2030	2040	2050	2060	2070	2080	2090	2100
World	GCAM4 - SSP4-26	Capacity|Electricity|Coal	GW		1225.958	1693.754	1228.808	1073.495	1071.386	931.691	628.734	233.791	50.596	8.569
World	IMAGE - SSP1-26	Capacity|Electricity|Coal	GW	967.114	1145.203	1358.195	1331.023	1204.439	921.282	741.316	878.261	891.990	806.116	591.645
World	MESSAGE-GLOBIOM - SSP2-26	Capacity|Electricity|Coal	GW	1208.581	1551.226	1615.484	1146.576	481.693	158.120	151.718	140.467	93.673	78.942	65.340
World	REMIND-MAGPIE - SSP5-26	Capacity|Electricity|Coal	GW	1134.000	1299.000	1297.000	705.500	187.100	2.116	1.084	0.601	0.309	0.104	0.077

© SSP Public Database (Version 2.0)
generated: 2019-08-13 16:13:27

Figure 37: Data viewer of the SSP database. Source: IIASA Website

I try to figure out the structure and functionality of the data viewer. It becomes clear to me while clicking through the elements that the content is structured from the top left to the bottom right, similarly to a book's page. Based on prior knowledge gathered through the field research, I click through the structure choosing the region 'world,' four different SSPs, a low scenario for future GHG emissions (RCP 2.6) and one particular simulated element (coal electricity). The diagram on the right now displays the data outputs of the simulations according to the choices made.

The x-axis represents time (years 2000 to 2100) and the y-axis the coal electricity capacity (gigawatt). I look at the diagram and try to interpret the form of the lines: If the world wants to limit global warming to 2 degrees (RCP 2.6), coal electricity should decrease as early as in the year 2020, independently of the prospect of the economy and other factors within the next years. I feel happy. After four years of my PhD research, I finally managed to grasp the basics of the SSPs to infer one major insight of the modeling work and to articulate it within this text. A little proud of this successful simulation of my informants' practices, I make a screenshot of the website and integrate it as an image above.

In a similar way as the models and code in chapter IV, we will now investigate how scientists make data and their information travel across contexts. On the one hand, this may help to obtain a better pragmatic understanding of 'research data.' On the other hand, it aims at characterizing the current transformation towards open data, infrastructures and services.

About data

There are not many terms in the contemporary world as omnipresent and problematic as 'data.' As Lisa Gitelman and Virginia Jackson have put it in the introduction of the anthology *'Raw Data' Is an Oxymoron*:

> Data are everywhere and piling up in dizzying amounts. Not too long ago storage and transmission media helped people grapple with kilobytes and megabytes, but today's databases and data backbones daily handle not just terabytes but petabytes of information, where *peta-* is a prefix which denotes the unfathomable quantity of a quadrillion, or a thousand trillion. (2013: 1)

However, data are not just *there,* they also *do* things with us and our world.

> Data are units or morsels of information that in aggregate form the bedrock of modern policy decisions by government and nongovernmental authorities. Data underlie the protocols of public health and medical practice, and data undergird the investment strategies and derivative instruments of finance capital. Data inform what we know about the universe, and they help indicate what is happening to the earth's climate. (ibid.)

This last reference to the Earth's climate points us to another crucial aspect of data. It is not only understood to give us information about the past and the present but also about our future. As Gitelman and Jackson quote from a famous IBM advertisement: "Our data isn't just telling us what's going on in the world, it's actually telling us where the world is going" (IBM cited in ibid.).

Ignoring the particular production context of this slogan (advertisement), this statement may actually be true and meaningful. In the particular case of climate research, the translation of GHGs and temperatures into handy datasets has not only enabled science to identify and prove a long-term trend toward global warming, it also provides a representational method of making statements about probable continuations of this trend in the future. Investigating 'data' in the context of climate impact research is interesting for a number of reasons. As a matter of fact, these researchers are engaging many things that have been problematized in other contexts: They have dealt almost exclusively with massive amounts of data ('Big Data'?) for a long time, engaged in complex analytic activities based on this data ('data science'?) and made predictions that have an impact on the lives of others (predictive analytics?). Without anticipating the following analysis in this chapter, we may argue that climate impact modelers do many of the things that are currently hyped and problematized elsewhere, but they do it a bit differently. Making use of a figure of speech introduced by Geoffrey Bowker, they seem to cook data with a lot of care (2005: 194). How may we characterize data? Rob Kitchin's seminal book *The Data Revolution* starts with the following preliminary description:

> Data are commonly understood to be the raw material produced by abstracting the world into categories, measures and other representational forms – numbers, characters, symbols, images, sounds, electromagnetic waves, bits – that constitute the building blocks from which information and knowledge are created. (2014b: 1)

The representative aspect of data for Kitchen is not always be explicit but may also be implied or derived:

> Data are usually representative in nature (e.g. measurements of a phenomena, such as a person's age, height, colour, blood pressure, opinion, habits, location, etc.), but can also be implied (e.g., through an absence rather than presence) or derived (e.g., data that are produced from other data, such as percentage change over time calculated by comparing data from two time periods) […]. (ibid.)

Data can either be recorded and stored in analog form or encoded as binary digits (bits). It may be categorized by form (qualitative or quantitative), structure (structured, semi-structured or unstructured), source (captured, derived, exhausted, transient), producer (primary, secondary, tertiary) and/or type (indexical, attribute, metadata) (ibid.: 4ff).

About research data

How can we get a grasp of scientific data and of the work it does? If we understand knowledge as situatively produced entities (Haraway 1988), it appears meaningless to characterize 'data' here in the abstract.

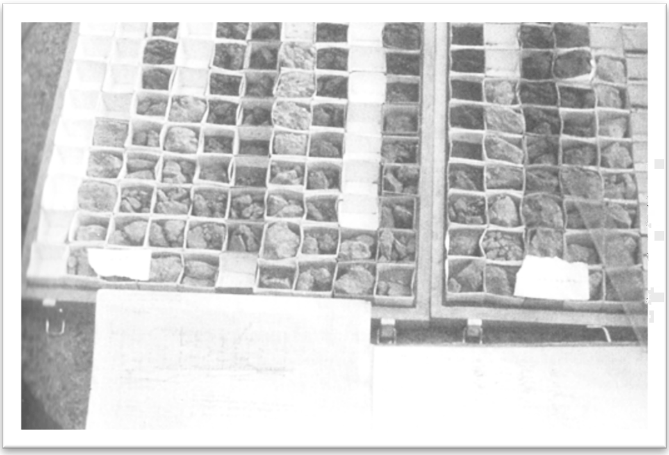


Figure 38: Soil stored in the pedocomparator transformed into an inscription. Source: Latour (1999a: 55)

Bruno Latour has addressed the construction and mobilization of scientific data in his article on *Circulating Reference* (1999a), tracing the transformation of the Amazon rainforest into botanic specimens, soil samples, tables, maps and, finally, into an academic publication. Building on this detailed description of transformative processes in science, we may ask: When does the forest cease to be forest and become 'data'? We should be careful of making this a categorical shift in science, but we can identify one instant that appears crucial in this becoming of data, namely, when soil samples stored and arranged in the instrument of the 'pedocomparator' are transformed into diagrammatic form on a piece of paper (see Figure 39): "We move now from the instrument to the diagram, from the hybrid earth/sign/drawer to paper" (ibid.: 54). This transformation is what Latour refers to as *inscription:*

> A general term that refers to all the types of transformations through which an entity becomes materialized into a sign, an archive, a document, a piece of paper, a trace. (Latour 1999b: 306)

In fact, 'inscription' for Latour is not only the process but also the resulting artifact:

> Usually, but not always inscriptions are two dimensional, superimposable, and combinable. They are always mobile that is, they allow new translations and articulations while keeping some types of relations intact. (ibid.: 306f)

This is where Latour equates 'inscription' with another of his concepts, the *immutable mobile*, "[...] a term that focuses on the movement of displacement and the contradictory requirements of the task" (ibid.:307). While we might contend that the 'inscription' and 'immutable mobile' oscillate well with concepts of 'data' (see Rheinberger's interpretation below), Latour's text is actually rather unspecific in his use of the term:

> [...] [A]n enormous pile of newspaper stuffed with plants brought back from the site and awaiting classification. The botanist has fallen behind. It is the same story in every laboratory. As soon as we go into the field or turn on an instrument, we find ourselves drowning in a sea of data. (I too have that problem, being incapable of saying all that can be said about a field trip that took only fifteen days.) Darwin moved out of his house soon after his voyage, pursued by treasure chests of data that ceaselessly arrived from the *Beagle.* Within the botanist's collection, the forest, reduced to its simplest expression, can quickly become as thick as the tangle of branches from which we started. (Latour 1999a: 39)

Therefore, 'data' is equally associated with specimens stuffed in newspapers, nonspecified material on Darwin's ship and, more broadly, with what we might call 'overwhelming impressions from the ethnographic field.' Latour shows a specific interest in the term 'data' only in one instance, highlighting its problematic etymology:

> In order for the botanical and pedological data to be superposed on the same diagram later, these two bodies of reference must be compatible. One should

> never speak of "data" – what is given – but rather of *sublata,* that is, of 'achievements.' (ibid.: 42)

This critique oscillates with arguments made by many others (Bowker 2005; Gitelman 2013; Kitchin 2014b; Leonelli 2015), which will be discussed more in detail further below. While Latour's article has not engaged in a fundamental characterization of data, it certainly influenced conceptualizations that followed. Hans-Jörg Rheinberger draws on Latour's arguments in his article *Infra-Experimentality*, translating the example of soil samples to genetics and the practice of genome sequencing:

> To stay with our molecular example, a next step consists in transforming the sequence gel into a chain of symbols standing for the four nucleic acid bases. With this visual display total abstraction is made not only from the particle from which the nucleic acid was extracted, but also from the test tube reaction in which it was differentially synthesized, and moreover from the gel and its material qualities in which the fragments were separated. (Rheinberger 211: 343)

In a similar way as Latour transforms soil samples in the pedocomparator to a map on paper, the sequence gel (trace) is transformed into a chain of symbols (data) containing the *information* for the expression of a protein. This is, according to Rheinberger, where "traces" become "data":

> The most important thing perhaps in such transitions is: the result of the experiment is brought into a form in which it can be *stored,* and consequently, *retrieved* as well. Much speaks for the assumption that the ability to be stored, that is, to be made *durable,* is the most important prerequisite for transforming *traces* into *data.* (ibid., emphasis in the original)

For Rheinberger, this is the shift in which immutable mobiles are born:

> Traces are not, but data are of the form of Latourian "immutable mobiles". Their relative immutability is a prerequisite for their mobility, their retrievability, their options for becoming re-enacted, and all the rest we associate with data and not with – usually precarious, bound-to-disappear – traces. (ibid., 344, emphasis in the original)

In this reading, data are synonymous with inscriptions and immutable mobiles. Data emerge in the moment when all material traces are exchanged against pure symbolic inscription. Both the articles by Rheinberger and Latour evoke the question to what extent the proposed characteristics of data, inscriptions and immutable mobiles are universal or specific to the sphere of research and the natural sciences in particular. This is especially the case for Latour, who has discussed the work of immutable mobiles more independently from science in his article on *Visualization and Cognition: Drawing Things Together* (Latour 1988). The text describes immutable mobiles in the form of cartographic inscriptions, which transformed the knowledge of power relations in colonial settings considerably. While maps may be produced by means of scientific instruments, the setting examined is clearly not one of science. Latour's work on immutable mobiles is generally so productive not only because it can stand for scientific knowledge production but also for knowledge production in general. The immutable mobile concept is itself an immutable mobile, traveling through the worlds of philosophy, the history of science and technology, and the sociology of knowledge. This consciously constructed vagueness of scale exists equally for the concept of Latour's *circulating reference:* "When immutable mobiles are cleverly aligned they produce the circulating reference" (Latour 1999b: 307) Latour is rather unspecific regarding in what reference frame this circulating reference is operating – as a philosophical category for

https://doi.org/10.14361/9783839452653-006 https://www.inlibra.com/de/agb - Open Access - (cc) BY

'sense-making in the world' or rather as a description of 'sense-making in science.'[74]

Data as relational property

Science studies scholar Sabina Leonelli builds on Rheinberger's and Latour's arguments but also criticizes their universal claims regarding data (inscriptions or immutable mobiles). The question for Leonelli what data is cannot be answered only by assessing its material qualities (mobility, stability across contexts) and degree of manipulation (inscription into symbolic form). Rather, she understands data as a purely relational property that can only be identified with reference to concrete research situations and the decisions and perceptions involved:

> A better option is give up altogether on a definition of data based on the degree to which they are manipulated, and focus instead on the relation between researchers' perceptions of what counts as data and the stages and contexts of investigation in which such perceptions emerge. (Leonelli 2015: 5)

In my opinion, this understanding of data has some strong argumentative points. Compared to Rheinberger's characterization, it diminishes the subliminal bias towards data produced within the natural sciences. While Rheinberger focuses on highly structured, numerical datasets, Leonelli's perspective may be better suited to incorporate the being of unstructured, heterogeneous data, for example, from ethnographic research or variations of web- and data-science: "Data can therefore

[74] Bruno Latour only makes explicitly clear that the description discards the perspective of sociology: "Of course had I not artificially severed the philosophy from the sociology, I would have to account for this division of labor between French and Brazilians, mestizos and Indians, and I would have to explain the male and female distributions of roles" (Latour 1999a: 44).

include experimental results as well as field observations, samples of organic materials, results of simulations and mathematical modeling, even specimens." (ibid.: 6)

Compared to Bruno Latour's immutable mobiles, Leonelli gives more weight to the prospective and perceptional aspects of data. From this perspective, the material form of artifacts is not the only constitutive feature of data. Equally, data must have been collected, stored and disseminated with the expectation of being used as evidence for knowledge claims. This does not necessarily mean that scientists know *how* the data might be used in the future (ibid.). This is an important point from an epistemological perspective and a critique of common understandings of data as "numbers, characters or images that designate an attribute of a phenomenon" (Royal Society cited in ibid.: 7). For her, this fundamental link between data and phenomena is not given. On the one hand, "researchers often produce data without knowing exactly which phenomenon they may document" (ibid.: 6). Researchers may produce data because they have access to particular instruments and they hope that it might later be helpful to identify new, unknown phenomena. On the other hand, the same data may act as evidence for a variety of phenomena, depending on the situational context (ibid.). However, similar to Latour and Rheinberger, Leonelli acknowledges the aspect of the *portability* of data as a precondition for its use as evidence:

> No intellectual achievement, no matter how revolutionary and well-justified, can be sanctioned as a contribution to scientific knowledge unless the individual concerned can express her ideas in a way that is intelligible to a community of peers, and can produce evidence that can be exhibited to others as corroborating her claims. […] If data are not portable, it is not possible to pass them around a group of individuals who can review their significance and bear witness to their scientific value. (ibid.: 7)

The characterization of data as 'portable objects' draws from common terminology in computer science, where software portability is understood as “a property of a program that can run on more than one kind of computer” (Downey 2012: 7). Equally, high-level languages (see Chapter IV) are portable, “meaning that they can run on different kinds of computers with few or no modifications. Low-level programs can run on only one kind of computer and have to be rewritten to run on another” (ibid.: 1). The term ‘data portability’ has only recently gained momentum in the context of the fight for digital rights. The recently established European General Data Protection Regulation, for example, includes the “right to data portability” (Article 20):

> The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided […].[75]

Data portability is closely related to the discourse of open data, which is introduced as follows.

Open data

In the past, access to valuable data and information has traditionally been restricted in some ways, for example, through financial, legal, organizational, technical and/or cognitive barriers. Accessing an academic publication, for example, might require the payment of a fee (financial) to a publisher in order to obtain the right (legal) to download an article. Equally, access to the publication might be restricted because it is only stored at one geographical location, in a state archive and in paper form (organizational). Vice versa, an exclusive availability online and in digital form creates new technical and cognitive barriers, as the discussion about the *digital divide* is showing (Norris 2001).

[75] https://gdpr-info.eu/art-20-gdpr/, retrieved on April 16, 2019.

Against this backdrop, the open data movement seeks to change this situation radically, making data potentially available to anyone.

As Rob Kitchin highlights, the movement is built on three principles: Openness, participation and collaboration. "Its aim is to democratize the ability to produce information and knowledge, rather than confining the power of data to its producers and those in a position to pay for access" (Kitchin 2014b: 48). On the one hand, attention has been focused on opening up public data emanating from state authorities and from research institutes, given that these have been funded by the public purse for the public's benefit. On the other hand, open data is also increasingly being pushed by the private industry, with anticipations of an innovative push through such practices. The open data community is interwoven with other movements fighting for the right to information, open knowledge, open-source software and open science. Within the last century, 'open data' has become increasingly popularized and mainstreamed through media campaigns (e.g. The Guardian's *Free Our Data*[76]), the call and endorsement of open data policies by inter- and supranational organizations (e.g. Organization for Economic Cooperation and Development[77] and the European Union[78]), national governments (e.g. Germany[79]) and municipal authorities (e.g. Berlin[80]) (ibid.).

Open Research Data

The idea of open public data had to be translated to the particularities of research practice to be meaningful for the case of science. This translation has recently given way to the *FAIR data principles*, a number of

[76] http://www.freeourdata.org.uk/, retrieved on April 5, 2019.

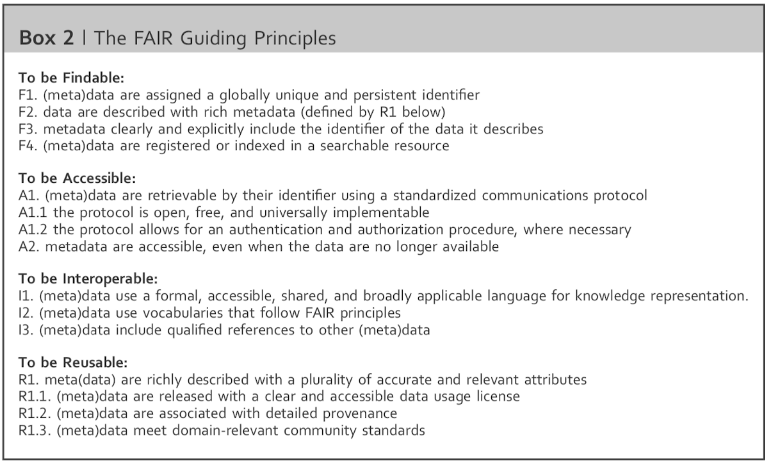
[77] https://data.oecd.org/, retrieved on April 5, 2019.

[78] https://data.europa.eu/euodp/en/data/, retrieved on April 5, 2019.

[79] https://www.govdata.de/, retrieved on April 5, 2019.

[80] https://daten.berlin.de/, retrieved on April 5, 2019.

guidelines supported and endorsed by various organizations in science and beyond:



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

> There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders – representing academia, industry, funding agencies, and scholarly publishers – have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. (Wilkinson et al. 2016: 1)

Table 2: The FAIR Guiding Principles.
Source: Wilkinson et al. (2016: 4)

'FAIR' denotates four different qualities of data: 'Findable,' 'Accessible,' 'Inter-Operable' and 'Re-usable' (see Table 2). Rather than discussing these principles in detail, we will now consider concrete open data practices in the field of climate impact research. As we will see, these practices oscillate with the aspects of knowledge mobilization discussed by Latour, Rheinberger and Leonelli.

Open data in action

We have already come across Tobias Geigers' work on extreme weather events and the simulation of their economic damage. Figure 39 shows a snapshot of an animation that Geiger uses to describe and explain his model and simulation, in this case, within my interview with him at the Potsdam Institute. The animation shows a dynamic simulation of damage from hurricanes within the territory of Bangladesh. It essentially works and aesthetically looks like a missile trajectory simulation.

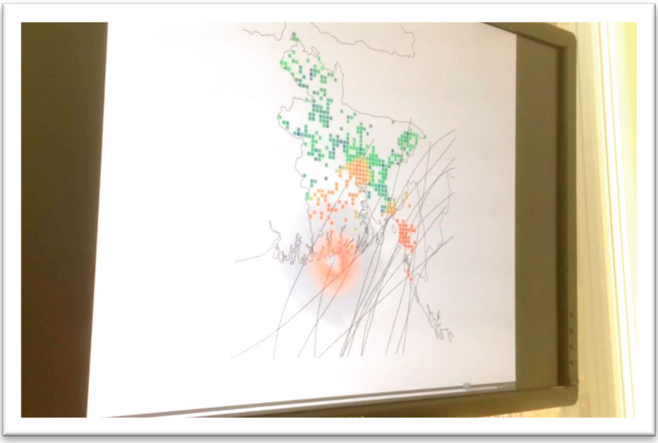


Figure 39: Animation of damage from hurricanes within the territory of Bangladesh. Source: https://vimeo.com/user49173690 by David Bresch, filmed during interview with Tobias Geiger

In the following, we will take a closer look at one particular dataset produced by Geiger and other scientists: Spatially-explicit Gross Cell Product (GCP) time series: past observations (1850–2000) harmonized with future projections according to the Shared Socioeconomic Pathways (2010–2100) (Geiger et al. 2017). The data consists of values for GDP in a temporal series spatial grid and temporal time series. The dataset has been used as input data for a number of scientific projects, such as the simulation of past and future damages of hurricanes modeled in/with CLIMADA (see Chapter IV), which is maintained by researchers at the Swiss ETHZ. It also builds on other work carried out at the Potsdam Institute, namely the SSPs as scenarios for future socioeconomic development. The dataset is stored on a server maintained by the shared library services of the research institutes on Telegrafenberg. It can be accessed via a library information sheet, similar to traditional academic publications. The description on the information sheet reads as follows:

> We here provide spatially-explicit economic time series for Gross Cell Product (GCP) with global coverage in 10-year increments between 1850 and 2100 with a spatial resolution of 5 arcmin. GCP is based on a statistical downscaling procedure that among other predictors uses national Gross Domestic Product (GDP) time series and gridded population estimates as input. Historical estimates until 2000 are harmonized with future socioeconomic projections from the Shared Socioeconomic Pathways (SSPs) according to SSP2 from 2010 onwards.
>
> We further provide a mapping file with identical spatial resolution to associate GCP values with specific countries. Based on this mapping we provide nationally aggregated GDP estimates between 1850–2100 in a separate csv-file.
>
> Additionally, we provide a mapping file with identical spatial resolution providing national assets-GDP ratios, that can be used to transform GCP to asset values based on 2016 estimates from Credit Suisse's Global Wealth Databook 2016.[81]

The terms of use of the dataset are specified as CC BY 4.0, a creative commons license allowing for sharing (copy and redistribute the material in any medium or format) and adapting (remix, transform and build upon the material) for any purpose, even commercial. However, the

[81] http://dataservices.gfz-potsdam.de/pik/showshort.php?id=escidoc:2740907, retrieved on April 5, 2019.

prospective user must give appropriate credit, provide a link to the license and indicate if changes were made. Equally, he/she is not allowed to apply legal terms or technological measures that legally restrict others from doing anything the license permits (Creative Commons 2019). In legal terms, this is the constitutional element to make this *open* data. The set includes four files (here with the description from the library page) in a zip-folder:

- GCP_PPP-2005_1850-2100.nc: GCP in 10-year increments between 1850 and 2100 with a resolution of 5 arcmin.
- National_GDP_PPP-2005_1850-2100.csv: nationally-aggregated GDP estimates (as used for GCP downscaling) in 10-year increments between 1850 and 2100.
- ISO-country-map.nc: Map for grid cell to ISO 3166 country code mapping with a resolution of 5 arcmin.
- GDP2Asset_converter_5arcmin.nc: Map for grid cell GDP to Asset mapping with a resolution of 5 arcmin based on 2016 estimates from Credit Suisse's Global Wealth Databook 2016.[82]

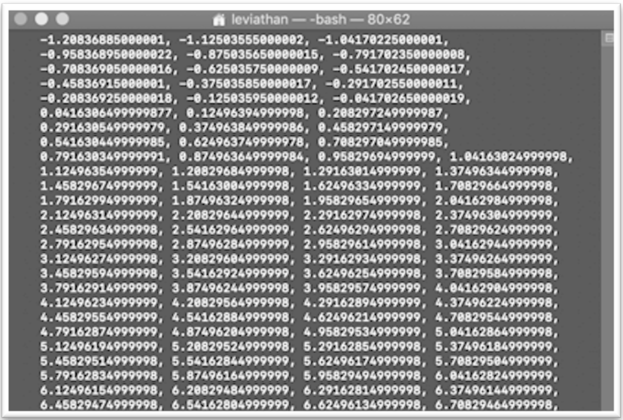
The four files are what my interview partners at PIK often referred to as 'raw data.' STS scholars, in contrast, have problematized this term on various occasions. The most prominent example is Geoffrey Bowker, who declared: "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care" (2005: 184) This idiom later inspired the anthology *Raw Data is an Oxymoron* edited by Lisa Gitelman (2013). The contributors in the anthology open various perspectives of the investigation of data. All articles are driven by the belief that we should refrain from considering 'data' in its etymological sense

[82] http://dataservices.gfz-potsdam.de/pik/showshort.php?id=escidoc:2740907, retrieved on April 5, 2019.

as given. Instead, we should carefully assess how data are 'cooked' within different social settings, circumstances and technological entanglements. In other words, data have history and this history matters. Against this background, a number of authors have introduced alternative terms to 'data,' considering their active generation. Exemplary are Bruno Latour's "sublata" (achievements) (1999a: 42) and Rob Kitchin's "capta" (taken):

> [...] what we understand as data are actually capta (derived from the Latin capere, meaning 'to take'); those units of data that have been selected and harvested from the sum of all potential data [...]. (2014b: 2)

As much as I acknowledge the critical consideration of data as 'given,' I would argue that the essence of data within scientific practice increasingly comes closer to its etymological origin. As Sabina Leonelli has shown, 'data' can only be considered as such when it is packaged for circulation. While I do not agree completely that this describes today's situation, the increasing mainstreaming of open data might soon make portability (or 'give-ability') conditional for 'data' to be considered as such in science.



```
    -1.20836885000001, -1.12503555000002, -1.04170225000001,
    -0.958368950000022, -0.875035650000015, -0.791702350000008,
    -0.708369050000016, -0.625035750000009, -0.541702450000017,
    -0.45836915000001, -0.375035850000017, -0.29170255000011,
    -0.208369250000018, -0.125035950000012, -0.041702650000019,
    0.0416306499999877, 0.12496394999998, 0.208297249999987,
    0.29163054999979, 0.374963849999986, 0.458297149999979,
    0.541630449999985, 0.624963749999978, 0.708297049999985,
    0.791630349999991, 0.874963649999984, 0.95829694999999, 1.04163024999998,
    1.12496354999999, 1.20829684999998, 1.29163014999999, 1.37496344999998,
    1.45829674999999, 1.54163004999998, 1.62496334999999, 1.70829664999998,
    1.79162994999999, 1.87496324999998, 1.95829654999999, 2.04162984999998,
    2.12496314999999, 2.20829644999999, 2.29162974999998, 2.37496304999999,
    2.45829634999998, 2.54162964999999, 2.62496294999998, 2.70829624999999,
    2.79162954999998, 2.87496284999999, 2.95829614999998, 3.04162944999999,
    3.12496274999998, 3.20829604999999, 3.29162934999998, 3.37496264999999,
    3.45829594999998, 3.54162924999999, 3.62496254999998, 3.70829584999998,
    3.79162914999999, 3.87496244999998, 3.95829574999999, 4.04162904999998,
    4.12496234999999, 4.20829564999998, 4.29162894999999, 4.37496224999998,
    4.45829554999999, 4.54162884999998, 4.62496214999999, 4.70829544999998,
    4.79162874999999, 4.87496204999998, 4.95829534999999, 5.04162864999998,
    5.12496194999999, 5.20829524999998, 5.29162854999998, 5.37496184999999,
    5.45829514999998, 5.54162844999999, 5.62496174999998, 5.70829504999999,
    5.79162834999998, 5.87496164999999, 5.95829494999998, 6.04162824999999,
    6.12496154999998, 6.20829484999999, 6.29162814999998, 6.37496144999999,
    6.45829474999998, 6.54162804999999, 6.62496134999998, 6.70829464999998,
```

Figure 40: Small fraction of the content
from file 'GCP_PPP-2005_1850-2100.nc' Source: My own screenshot

In my opinion, the term 'raw data' is actually less problematic than 'data' in general. At least in the case of computational science, it has a distinct meaning and justification in practice. 'Raw' refers to the data that is machine-readable but not manipulated to enable human cognition. In everyday language, scientists may point to columns of numbers when referring to 'raw data,' as shown in Figure 40.

The snapshot depicts a small fraction of the content from file 'GCP_PPP-2005_1850-2100.nc' as retrieved by the netCDF reader (,ncdump –ct' command) and visually represented by my MacOS terminal. In a strict sense, the content represented through the snapshot is no longer 'raw data' but processed through algorithms to make it visible and cognizable for me as a human being. The raw version of the dataset cannot be shown as it is invisibly stored in a digital database[83] on my laptop's hard drive. Accordingly, when scientists talk about 'raw data,' they might point to visible columns of numbers, but they are understood as an index for the invisible information stored in the database at stake.

The use of the term 'raw data' is also rooted in principles of computational scientific practice. The scientists in my field always seek the maximum proximity to 'raw data' possible. Practically, this means that they prefer to work within text terminals or consoles, write their own software wherever possible and are generally skeptical toward software written by others, visual representations instead of numbers, and comprehensive proprietary tools. As a matter of fact, the frequent use of the term 'tool' in my interviews with climate impact modelers oscillate with Jörg Rheinberger's characterization of *technical objects* (1997) and its differentiation against *epistemic things*, which has also been discussed for the case of simulation models by Mikaela Sundberg (2008). In my

[83] Marcus Burkhardt has shown that the term database is ambiguous in its meaning and use. It may refer equally to collections in general and collections of digital information in particular, i.e. the technologies that process structured collections of machine-readable information (2015: 131).

interviews, my informants referred to 'tools,' meaning software developed by others and beyond the control of the researcher:

> So, GRASS and Mapwindows, these are [...] tools that are used in this field. (Interview Willkomm)

> If you click here, [...] it's their tool that does everything. And they calculate everything internally, which then enables you to navigate here. (Interview Rechstein)

In contrast to epistemic things, such as 'question-generating machines' (e.g. the simulation models), tools are understood as technical objects or 'answering machines' (Rheinberger 1997; Sundberg 2008). They are problem solvers, not knowledge producers. As scientists are aware of the frictions between these two functions (Knorr-Cetina 2003), they traditionally refrain from using technical constellations with a (perceived) high epistemic opacity (more on these aspects in Chapter IV).

The differentiation between objects within and beyond the control of the scientist is also valid for the case of data, which bows down to the distinction between 'primary' and 'secondary data.' Primary data refers to data generated by the researchers themselves, making use of their own instruments, according to their proper research design and methodology (Kitchin 2014b: 7). By contrast, researchers often make use of *secondary data* generated and provided by others, possibly with very different instruments and research designs. In the case of the modeling work described here, for example, the researchers have used externally generated data to produce their own global dataset ('primary data') of historic and future GDP. This secondary data comes from heterogeneous resources and actors, such as the Madison Project

Database[84] (University of Groningen) and the Global Wealth Databook[85] (CS Credit Suisse bank). This fluidity of (i.e. 'open') data between modeling groups of different institutions, scientific fields (physics, economics) and production contexts (scientific, commercial) creates new opportunities but also challenges of trust, which are addressed by numerous strategies of standardization, documentation and evaluation.

The data files in our example, for instance, are encoded in widely standardized formats, such as NetCDF and comma-separated values (commonly known as a csv). The dataset is equipped with a digital object identifier (seen as DOI in bibliographies[86]), a format for unique resource identification. Kitchin highlights that such *indexical data* enable identification and linking;

> [...] indexical data are important because they enable large amounts of non-indexical data to be bound together and tracked through shared identifiers, and enable discrimination, combination, disaggregation and re-aggregation, searching and other forms of processing and analysis. [...] Indexical data are becoming increasingly common and granular, escalating the relationality of datasets. (Kitchin 2014b: 8)

The dataset also includes four metadata files in xml (eXtensible Markup Language) format, according to the standards iso19115, datasite, dif and escidoc. *Metadata* are essentially data about data.

> Metadata can either refer to the data content or the whole dataset. Metadata about the content includes the names and descriptions of specific fields (e.g., the column headers in a spreadsheet) and data definitions. These metadata help a user of a dataset to understand its composition and how it should be

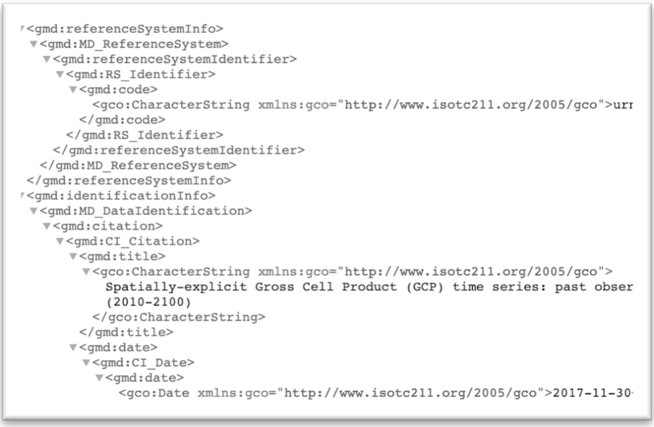
[84] https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2018, retrieved on December 10, 2020.

[85] https://www.credit-suisse.com/about-us/en/reports-research/global-wealth-report.html, retrieved on December 10, 2020.

[86] http://doi.org/10.5880/pik.2017.011, retrieved on December 10, 2020.

> used and interpreted, and facilitates the conjoining of datasets, interoperability and discoverability, and to judge their provenance and lineage. (ibid.: 8f)

The formatting of the files in the markup language XML ensures that the information stored is both human- and machine-readable. Figure 41 depicts a snapshot of the iso19115-file[87] in its "raw" machine-readable version for illustrative purposes.



```
<gmd:referenceSystemInfo>
 ▼<gmd:MD_ReferenceSystem>
   ▼<gmd:referenceSystemIdentifier>
     ▼<gmd:RS_Identifier>
       ▼<gmd:code>
          <gco:CharacterString xmlns:gco="http://www.isotc211.org/2005/gco">urr
        </gmd:code>
      </gmd:RS_Identifier>
    </gmd:referenceSystemIdentifier>
  </gmd:MD_ReferenceSystem>
</gmd:referenceSystemInfo>
<gmd:identificationInfo>
 ▼<gmd:MD_DataIdentification>
   ▼<gmd:citation>
     ▼<gmd:CI_Citation>
       ▼<gmd:title>
         ▼<gco:CharacterString xmlns:gco="http://www.isotc211.org/2005/gco">
            Spatially-explicit Gross Cell Product (GCP) time series: past obser
            (2010-2100)
          </gco:CharacterString>
        </gmd:title>
       ▼<gmd:date>
         ▼<gmd:CI_Date>
           ▼<gmd:date>
              <gco:Date xmlns:gco="http://www.isotc211.org/2005/gco">2017-11-30
```

Figure 41: Snapshot of a metadata file according to iso19115.
Source: My own screenshot

In addition, a hyperlinked bibliography cites other datasets and scientific publications, which are related to the present set.

Linked data

There has been a shift in the understanding of open data from a category of human-human and human-machine interaction to a category of

[87] https://dataservices.gfz-potsdam.de/pik/download.php?item=escidoc-2740907&mdrecord=iso19115.

machine-machine readability as I could witness in my fieldwork at the PIK and during an Open Science fellowship at the Wikimedia Foundation. As an example of the latter, we can quote from the FAIR data principles, which highlight the significance of "computational stakeholders":

> Humans, however, are not the only critical stakeholders in the milieu of scientific data. Similar problems are encountered by the applications and computational agents that we task to undertake data retrieval and analysis on our behalf. These 'computational stakeholders' are increasingly relevant, and demand as much, or more, attention as their importance grows. (Wilkinson et al. 2016: 2)

This coincides with a blending of the open data discourse with the idea of *linked data,* which is the idea to transform the internet from a 'web of documents' to a 'web of data':

> Such a vision recognises that all of the information shared on the Web contains a rich diversity of data – names, addresses, product details, facts, figures, and so on. However, these data are not necessarily formally identified as such, nor are they formally structured in such a way as to be easily harvested and used. (Kitchin 2014b: 52)

Tim Berners-Lee (2009) mentions four rules of behavior for such linked data in a semantic, machine-readable web: The first is to identify things with Unified Resource Identifiers (URIs): "If it doesn't use the universal URI set of symbols, we don't call it Semantic Web" (ibid.) The second rule is to use a particular type of such identifiers, namely Hypertext Transfer Protocol (HTTP) URIs, which works well with the inherent structure of the internet. The third rule is that one should provide meta-information on the web against a URI. Berners-Lee mentions a number of standards (RDF, XML) that can be searched by dedicated query languages (SPARQL). The fourth important rule is to produce links to these URIs elsewhere, "which is necessary to connect the data we have into a web, a serious, unbounded web in which one can find all kinds of things, just as on the hypertext web we have managed to build" (ibid.). A prominent example of a linked data project is the open knowledge base *Wikidata,* a machine-readable supplement and further development of the ideas around Wikipedia. In contrast to Wikipedia, the information (data) on Wikidata is highly structured in the ways described above.

As an example, we can take a look at the Wikidata entry for 'data'[88] (see table 3). Items are uniquely identified by a 'Q' followed by a number, in this case 'Q42848.' Statements describe detailed characteristics of an item and consist of a property (e.g. instance of) and a value (e.g. abstract object). Properties are identified by a 'P' followed by a number, such as 'subclass of (P279).'

Item	**data (Q42848)**
description	facts represented for handling
statements	
instance of (P31)	abstract object (Q7184903)
subclass of (P279)	information (Q11028)
part of (P361)	data base (Q59138835)
topic's main category (P910)	category: Data Q6641340
different from (P1889)	knowledge (Q9081)
identifiers	
JSTOR topic ID (P3827)	scientific data
(…)	

Table 2: Wikidata entry for 'data' (Q42848). Source: My own table

[88] https://www.wikidata.org/wiki/Q42848, retrieved on December 10, 2020.

By means of Wikimedia's query service,[89] one can search through the Wikidata knowledge base and conduct comprehensive activities of data analysis.

In the context of our study, the Wikidata project may serve as an example for the sociotechnical imaginary (Jasanoff/Kim 2015) of a web of data. As a matter of fact, it is an imaginary not only supported by civil society organizations, such as Wikimedia, but also by powerful actors of the knowledge economy. From the beginning, Wikidata had been co-funded by Google and the Allen Institute for Artificial Intelligence, an organization established by Microsoft co-founder Paul Allen.[90] These actors are highly interested in the mainstreaming of linked open data, which promises numerous market opportunities. In this perspective, Google recently launched its own dataset research engine,[91] which enables one to search the web for structured datasets, and research data specifically. Within a web of data, the spheres of science and private business become ever more permeable. This permeability of data and its production contexts creates new challenges for scientific practice, as the following interview excerpt shows:

> We will derive a damage function, with data from the reinsurance company Munich Re, and also from SwissRe. And the data is not publicly available. And we use it to derive these damage functions, which can then be used by anyone. But in order to be able to reproduce the loss function, you have to get this data from Munich Re yourself. [...] There are a few other datasets that are publicly available for these damages, but they have some kind of spatial and also a temporary bias. (Interview Geiger, translated by the author)

[89] https://query.wikidata.org/, retrieved on April 3, 2019.

[90] http://tcrn.ch/H0aO9U, retrieved on April 3, 2019.

[91] https://toolbox.google.com/datasetsearch, retrieved on April 3, 2019

The impossibility of reproducing the scientific process due to the use of proprietary information as input data for simulations is problematic from a perspective of scientific verifiability and political accountability. Moreover, it creates a shift in the distribution of work between scientific and other actors, which can also be a trigger for controversy:

> Yes, that is in a way an outsourcing of their own research activities. [...] So, there are different opinions on that. The Munich Re clearly says that they are very interested in doing research with it. Because they benefit from it. Swiss Re then seems to be a bit more reserved [...]. In addition, there are others, pseudo-insurers, or over-insurers in the United States, who simply earn a lot of money with these data, by selling them to the insurance industry. They don't make them freely available. (Interview Geiger)

Of course, the protection of scientific independency from economic interests is not a new issue, but it gains new relevance in the context of 'climate services' (Krauss/von Storch 2012; Vaughan/Dessai 2014), and data-driven and sustainability research in general.

Open data infrastructure

As we have already seen in some of the examples discussed above, open data is not just a matter of appropriate licensing but requires the setup of comprehensive new infrastructures. In Chapter I, I briefly mentioned the repurposing of the existing infrastructural base at the Science Park Albert Einstein for contemporary techno-scientific challenges. The most virulent example is the repurposing of the previous Geodetic Observatory into the library shared by all institutes of the science park. The main reading room of the library is located at the former 'great instrument hall.' The 'hall of the pendulum,' which had hosted the consequential geodetic experiments by Friedrich Jakob Kühnen and Philipp Furtwängler (Kühnen/Furtwängler 1906; Reicheneder 1959), has now been converted into a museum of geodetic instruments (see Figure 42).

Figure 42: Museum of geodetic instruments at the library of Telegrafenberg. Source: GFZ (2012)

In fact, the library is distributed in several buildings on the hill, including the PIK headquarters at Michelson House. This productive collaboration between the four institutes on the hill is not self-evident, given the extraordinary ambience of competition between the organizations. As a former employee of the GFZ IT services highlights:

> This becomes increasingly relevant, considering that the library also serves as data repository for so-called 'gray literature,' which includes software and data.
> (Interview Gephardt, translated by the author)

Co-funded by the institutes on the hill, the library is currently investing many resources to build such data infrastructure. This includes the hosting and documentation of datasets such as the one discussed above and a sophisticated system and interface to make the data accessible. One can browse a global Leaflet map (see Chapter VI), for example, to search for datasets that address a specific region. Figure 43 shows such a spatial representation of data available for the PIK's dataset 'Simulation Data from Water (regional) Sector.'[92]

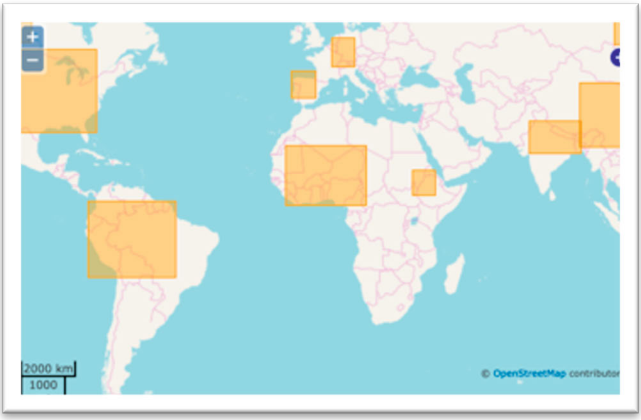


Figure 43: Leaflet map of the spatialities of a dataset.
Source: GFZ Data Services

We can see that 'open data' is not only a matter of saving datasets in an appropriate format or adding meta-data to it, but also involves the development of a comprehensive new infrastructure, with its own entanglement of physical, symbolic and – not to forget – human elements.

Open data is work

This last aspect – human engagement – is often forgotten in the discussion about open data. In my interview with IT infrastructure expert Paul Gephardt, he highlights that open data infrastructures will have considerable consequences for the life realities of the scientists:

[92] http://dataservices.gfz-potsdam.de/pik/showshort.php?id=escidoc:2959917, retrieved on April 3, 2019.

> […] research data management plans will be necessary, the provision of safe data repositories, data must be labelled with Digital Object Identifiers. And to implement the entire transformation process towards an open science paradigm into reality.
> (Interview Gephardt)

It seems useful to reconsider open data as an objectual category and investigate more thoroughly what 'opening data' means as a practice. Michael Gurstein, among others, has criticized this sole focus on objectual characteristics of open data:

> As an object or thing the attributes and characteristics of the open data are more or less fixed once made available to the end user/consumer. As well, the determination of the attributes or characteristics of the data (what the open data "is") as seen/obtained by the end user is solely at the discretion of the producer and are uniform and stable as between end users. (2013)

Instead of such an understanding of open data as objects and products, Gurstein proposes a focus on open data as a service:

> But why shouldn't we think of 'open data' as a 'service' where the open data rather than being characterized by its 'thingness' or its unchangeable quality as a 'product', can be understood as an on-going interactive and iterative process of co-creation between the data supplier and the end-user […]. (ibid.)

In so doing, one could put more emphasis on 'opening' as a transitive and interactive process, an interaction and relationship between suppliers and users. For Gurstein, this reconceptualization would have consequences for the way in which open data is funded, managed and made available. It would require a review of the relationship between the open data discourse and neoliberal agendas marketizing public services.

> […] if one treats open data simply or exclusively as a thing or commodity then it is available solely as a product for purchase and use through the market place–where of course, market principles dominate and where for example, those with the most resources are able to command and control and thus precipitate the supply of the product i.e. the open data. (ibid.).

This oscillates with the 'hijacking' of the open data discourse by the big players of the knowledge economy discussed above. In an earlier article written in 2011, Gurstein proposes a number of necessary elements enabling a more effective and inclusive use of open data, thereby reducing the "data divide." These include:

- available telecommunications/Internet access;
- having access to machines/computers/software;
- having sufficient knowledge/skill to use the software required for the analyses;
- having the data available in a format to allow for effective use at a variety of levels of linguistic and computer literacy;
- sufficient knowledge and skill to see what data uses make sense (and which do not) and to add local value;
- having supportive individual or community resources sufficient for translating data into activities for local benefit; and
- the required financing, legal, regulatory or policy regime, required to enable the use to which the data would be put.

(summarized from Gurstein 2011: 5f)

Within the years, general agreement on these challenges and needs have enabled large resources to build up open data infrastructures and services. Examples are the Data Services at the Science Park Albert

Einstein discussed, and much more extensive initiatives, such as the National Oceanographic and Atmospheric Agency (NOAA) Data Discovery Portal[93] in the United States or Worldbank Open Data.[94] The strategic importance of such infrastructures and data centers became clear at the time of the inauguration of US President Donald Trump. Due to his regular statements describing climate change as a hoax, climate scientists and civil society organizations feared that the new president would shut down the research programs on climate change, leading to a deletion of climate-related data stored on government servers, for example, data hosted by the NOAA, the Environmental Protection Agency and the White House. Based on such fears of data demolition, scientists from Penn and other Universities organized DataRefuge, a large-scale data migration of US government data to servers in Canada. While the scenario of data demolition has not yet materialized,[95] the episode shows how open data is dependent on working infrastructures and institutional support. As a matter of fact, the main value of the DataRefuge project might not have been to copy datasets from one server to the other. Rather, the project triggered activities of infrastructural inversion: While environmental data has formerly been taken for granted, the wide-ranging discussion around DataRefuge generated a variety of initiatives surfacing datasets, considering their relevance and characteristics, and making them more effectively available. The chief data officer of NOAA, Edward J. Kearns, for example, published a comprehensive statement, where he mapped out the whole data infrastructure of the institution and addressed the public worries considering politically motivated data deletion:

[93] https://data.noaa.gov/datasetsearch/, retrieved on April 3, 2019.

[94] https://data.worldbank.org/, retrieved on April 3, 2019.

[95] https://sunlightfoundation.com/tracking-u-s-government-data-removed-from-the-internet-during-the-trump-administration/, retrieved on April 2, 2019.

> I am sometimes asked if NOAA's data in its archives can be easily deleted. No they can't, since data may not be removed without significant effort and public deliberation. It is also unlawful to tamper, damage, delete, vandalize, or in any way alter formal federal records, including NOAA's environmental data and its archives.[96]

Figure 44: DataRefuge logo.
Source: www.datarefuge.org

Kearns also criticized an exaggerated focus on the datasets as objects, ignoring the crucial contribution of open data work within research infrastructures:

> The value of NOAA's data archives include not just the simple existence of the data themselves, but the continuous investment of NOAA's experts' efforts towards the sustained quality and usability of the data. The integrity and accuracy of data that are stored on non-federal system and are not stewarded by NOAA's scientists cannot always be easily verified beyond file-level distribution. NOAA is currently exploring best practices and technologies that may allow the authentication of its data throughout the wider data ecosystem, and welcomes interested parties in academia and industry to join in this exploration.[97]

Notwithstanding these insurances by Kearns, the activists of DataRefuge continued to move datasets from governmental servers, which turned out to be extensive and tricky work. The activists had to invent new methods of identifying, understanding, copying, ordering, monitoring and making datasets accessible. Given the effort and inventive-

[96] https://libraries.network/blog/2017/4/30/on-the-preservation-of-and-access-to-noaas-open-data, retrieved on April 2, 2019.

[97] https://libraries.network/blog/2017/4/30/on-the-preservation-of-and-access-to-noaas-open-data, retrieved on April 2, 2019.

ness, they came up with a new job description for these tasks: The 'data baggers' write custom scripts to scrape complicated datasets from distributed sources and patched-together federal websites. A coverage of the *Wired* magazine shows that DataRefuge led to a veritable imagination of a parallel infrastructure monitoring research infrastructure:

> […] two dozen or so of the most advanced software builders gathered around whiteboards, sketching out tools they'll need. They worked out filters to separate mundane updates from major shake-ups, and explored blockchain-like systems to build auditable ledgers of alterations. Basically it's an issue of what engineers call version control – how do you know if something has changed? How do you know if you have the latest? How do you keep track of the old stuff? […] DataRefuge and EDGI understand that they need to be monitoring those changes and deletions. That's more work than a human could do. So they're building software that can do it automatically.[98]

Open research data are not as fluid as one could think, but rather a sticky matter. As the example of DataRefuge shows, seamless data portability imagined by the open data discourse is a tricky object of desire. The more one wants to mobilize datasets and reduce the seamfulness in systems (Vertesi 2014), one is obliged to invest in the construction of a fluidifying data infrastructure.

Infra-worlds of knowledge

Based on the preceding discussions, I would like to come back to the literature discussed at the beginning of this chapter and propose some reconsiderations of the status of research data in contemporary computation science. As Rheinberger highlights, DNA sequencing has been delegated to automated analyzers for several decades now. The field of bioinformatics has been developed with its own set of methods and

[98] https://www.wired.com/2017/02/diehard-coders-just-saved-nasas-earth-science-data/, retrieved on April 2, 2019.

infrastructures to domesticate this plethora of data. For Rheinberger, this represents a new phase in the relationship between (molecular life) science and information. While it was formerly the discursive and conceptual aspects of information that have been prominent in molecular genetics, bioinformatics has shifted this focus to the sphere of the infrastructural:

> Data have become a resource, rather than a result in the world of infra-experimentality, produced on an industrial scale and made intelligible only in the context of appropriate software. The research technologies in the space between the knower and the to-be-known have entered the stage of a second order mediation. The data, mediators between traces and technophenomena, have proliferated and created a world of their own. (Rheinberger 2011: 346)

These "infra-worlds of knowledge" (ibid.) in the field of simulation modeling do not only gain momentum in daily scientific practice but also as a matter of scientific reputation. *Inter alia,* this valuation is demonstrated by the rise of the *data publication.* While there is significant debate around the formats, processes and terminology of this new publication format, the general purpose is to "bring datasets into the scholarly record as first class research products (validated, preserved, cited, and credited)" (Kratz/Strasser 2014: 1). The format promises deliverables for various actors, including scientists, journal editors, publishers, data centers, the scientific community, funding agencies, governments and society as a whole. According to a conceptual paper from 2009 (Costello 2009: 420), benefits include additional publications and higher citation rates for individual researchers, possibilities for verification and accountability, greater valuation of data and data producers, higher financial return on research investments and, simply put, "better science."

We can illustrate the concrete realization of the format in scientific practice with the example from the PIK relating to the dataset discussed above: *Continuous national gross domestic product (GDP) time series for*

195 countries: Past observations (1850–2005) harmonized with future projections according to the Shared Socio-economic Pathways (2006–2100) (Geiger 2018). The data publication appears in the journal *Earth System Science Data*, which presents itself as "an international, interdisciplinary journal for the publication of articles on original research data (sets), furthering the reuse of high-quality data of benefit to Earth system sciences."[99] The publication begins with a discussion of the underlying metric for the dataset, the GDP. The author describes the GDP's role as a standard indicator for assessing a nation's development and discusses the criticisms regarding its representational features regarding growth, development, and welfare and well-being. Despite these limitations and because of a lack of alternatives, "GDP has proven to be a useful measure to track the evolution of economic development within or across nations" (ibid.: 487).

The publication then describes input data and the methods used to create the dataset in question: A continuous and consistent GDP time series for 195 countries. Input data includes the Penn World Table, the Maddison Project Database, World Development Indicators, the History Database of the Global Environment and future projections from the SSPs. The discussion in the methodological section addresses ways to deal with missing data and interpolation:

> As a first step we populate all missing data points in 1850, the initial year of our data product, by linear interpolation between the last available data point before 1850 and the first one after 1850, ensuring that it is not more distant in time than 1870. Next, and if available, we generate annual data by linear interpolation of data points between 1850, 1860, and 1870. These preparatory steps reduce the missing value fraction from 51.7 to 48.5 %. (ibid.: 850)

[99] https://earth-system-science-data.net/, retrieved on April 2, 2019.

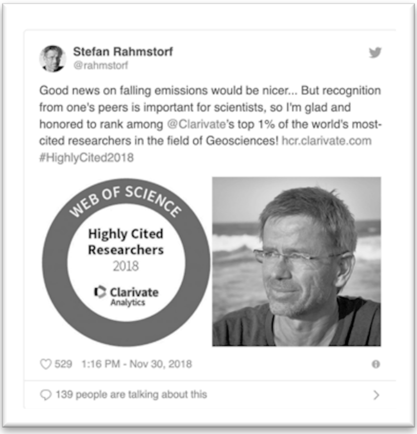
The text focuses specifically on calculation issues for particular regions of the world. The Balkans, for example, create considerable challenges for calculability due to frequent territorial alterations and missing data in times of conflict. Equally, data quality for the African continent is extraordinarily low:

> The MPD [Maddison Project Database] contains only six countries with income data prior to 1950: Egypt, Tunisia, Morocco, Algeria, and South Africa/Cape Colony (all since 1820), and Ghana (since 1870). Therefore, the African total (AT) population-weighted average income prior to 1950 is only defined by six countries. For historic and geographic reasons, we assume that those countries define the upper income limit when extrapolating the remaining countries back in time. (ibid.)

From a representational point of view, data for this region is generated basically by the author alone and only rarely represents original data produced 'on the ground.' After the explanations of the method and process, the publication describes the resulting datasets: "We provide three different primary data sets, a data description file, and two supplementary data sets in the online archive at https://doi.org/10.5880/pik.2017.003" (ibid.: 854). In the last section of the paper, the author makes an assessment of the quality of the dataset:

> While rather exhaustive data exist for Western European countries, these limitations might be less of a problem than for most African countries. As a consequence, one should treat the data with care and allow for uncertainties, in particular where data coverage is limited or almost non-existent. (ibid.)

"Treat the data with care" refers to the prospective users of the dataset. This reuse by scientific or nonscientific actors is the primary objective of the data publication (understood as the practice and resulting set of artifacts). However, as has been mentioned earlier, it also becomes a matter of reputation to make datasets public. Thanks to the format of the data publication, datasets become citable within global citation databases, such as the Web of Science,[100] Scopus and Google Scholar. These bibliometric and scientometric evaluation schemes of science (Sengupta, 2009) increasingly dominate the reputation systems of the researchers and their institutes. The relevance of these schemes can be illustrated by a tweet of the PIK researcher Stefan Rahmstorf, representing his status in the "top 1 % of the world's most-cited researchers in the field of Geosciences" on the Web of Science (see Figure 45). The tweet[101] has been pinned to Rahmstorf's feed for almost a year.



Stefan Rahmstorf
@rahmstorf

Good news on falling emissions would be nicer... But recognition from one's peers is important for scientists, so I'm glad and honored to rank among @Clarivate's top 1% of the world's most-cited researchers in the field of Geosciences! hcr.clarivate.com #HighlyCited2018

♡ 529 1:16 PM - Nov 30, 2018

139 people are talking about this

Figure 45: Pinned tweet by PIK researcher Stefan Rahmstorf. Source: Twitter

Against this background, data publication becomes increasingly important as a device to perform in these reputation schemes. By producing and publishing data with an open license requiring attribution (e.g. CC BY), one makes sure that all future computational work involving the datasets generates a citation in bibliometric schemes.

[100] https://clarivate.com/webofsciencegroup/solutions/web-of-science/, retrieved on April 2, 2019.

[101] Tweet posted on November 30, 2019, retrieved on September 3, 2019, via https://twitter.com/rahmstorf/status/1068463676628312064.

I would argue that this has significant consequences for the scientific practices in computational science. As the example of the GDP dataset shows, the aim of published data is not to create a new epistemic thing but a technical object, in the sense described by Jörg Rheinberger (1997). It is a forward-looking scientific practice aiming at the production of a robust base for further calculations. Literally and metaphorically, the dataset constitutes a spatiotemporal 'grid' that will drive further simulations. Against this background, the choice to base this grid on GDP is only logical. It makes sure that the dataset is not only used within the world of climate research but can potentially be used by virtually anyone simulating economic and social development in the world. The compatibility of the time series with future scenarios of socioeconomic development generated at the PIK (i.e. SSPs) makes it possible to use the dataset as a base for all sorts of predictions for the 21st century. More than 'facts,' 'information' or 'evidence,' these datasets constitute infrastructural elements for all sorts of prospective knowledge claims to be generated by scientific and nonscientific actors. Against this background, I would refer to them as *infrastructural data,* as they can build a fundament for multiple future work.

In my interviews, climate modelers often used the terms of 'drivers' and 'drivers of the future.' 'Drivers' refer to computer models that provide the conditions for other models. Such drivers are models that drive other models further down the model chain.

> This means that the driving model is always a global model, and the regional model then provides much more precise climate information for Greater Europe, due to the fact that the topography is resolved much more finely. (Interview Hauser, translated by the author)

On a material level, it is not the model that drives future calculations but its outputs in the form of datasets, for example, a time-series of GDP, temperature or CO_2data. The focus of work in the field of climate impact research is increasingly shifting to data(sets). While climate modelers tend to dislike the characterization of the 'data scientist,' their field is surely subject to datafication on various levels. We might speak of a datafication of computational science in this context, rather than of data-science. Nevertheless, the Potsdam Institute recently used the connotation of the 'data scientist' for the first time in a job advertisement. The PIK is also a member institution of *GeoX Data Science*,[102] an innovation program funded by the Telegrafenberg institutions and seven other geoscientific facilities in the Berlin-Brandenburg area.

A pragmatist typology of data

We should reconsider the status of categories such as 'data,' 'research data' and 'evidence' carefully. Being conscious of the limits of comparison, we can draw some parallels between Rheinberger's observations in the field of bioinformatics and climate-related simulation modeling. Regarding the example of the *Spatially-explicit Gross Cell Product (GCP) time series* discussed above, we have seen that the relevant dataset includes a variety of different 'data,' which may individually be labeled as primary data, secondary data, tertiary data, indexical data, metadata, and so on. The label does not characterize any essential (material) qualities of the data but its relational entanglement and situated context of use. As Leonelli highlights: "Depending on what uses the data are eventually put to, and by whom, those modifications may well prove as relevant to making data into valuable evidence as the efforts of the original data producer" (2015: 8). While I agree with this argument, we may want to reconsider the predominant role that Leonelli gives to the category of scientific data as 'prospective evidence.' Instead, I propose a categorization of data according to its concrete use in research practice, which goes well beyond its prospective use as scientific evidence.

[102] www.geo-x.net/en/, retrieved on April 3, 2019.

Pragmatist categories of research data:

1. Evidential data

Evidential data serve as proof of a certain attribute of a phenomenon and establish scientific facts. In climate impact research, evidential data may represent predictions regarding future developments, evaluations of future risk or correlations between such variables.

2. Infrastructural data

Infrastructural data are not produced with the aim of providing evidence but to provide a stable foundation for manipulations of other data. They are evaluated, refined and optimized to serve these purposes. The becoming of infrastructural data requires cumbersome work of testing, standardization and evaluation. This means that the becoming of infrastructural data takes time and involves cooperative efforts in communities. Examples in the context of climate simulation modeling are standardized geospatial databases and time series (e.g. the series of GDP values discussed earlier).

3. Resourceful data

In this context, data serves as a resource for the identification of patterns, thereby enabling the extraction of further information and knowledge. It is the role of data within practices referred to as 'data-driven research' or 'data-science.' A popular example of resourceful data is ImageNet,[103] a large image database currently containing more than 14 million images, classified along object properties. ImageNet has

[103] Description from the ImageNet website: "ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures." Retrieved on September 23, 2019, via http://www.image-net.org/.

been used most prominently as a training dataset for convolutional neural networks since the 2012. Another example is the use of social media data in the context of *digital methods* research (Rogers 2013; 2015) and its dedicated software tools (e.g. Borra/Rieder 2014; Rieder 2015). As a matter of fact, researchers involved in the Digital Methods Initiative have often used a 'climate change dataset' within their studies, given its quality as a dataset. We can exemplify this with a text passage from Noortje Marres and Carolin Gerlitz's article on *Interface Methods:*

> For our analysis of 'happening content,' we decide to focus on a fairly general issue term, namely *climate change*, and include in our data set all Tweets using this term for a period of almost three months – from March 1st, 2012 to June 15, 2012, adding up to a total of 204795 tweets, a workable, medium-sized data set. (2015: 17)

In this context, 'climate change data on Twitter' was not primarily chosen in order to make statements about public debates on climate change but due to its quality as a "workable, medium-sized data set."

4. Communicational data

Communicational data enable the mobilization and resulting portability of evidential, infrastructural or resourceful data. In so doing, communicational data permits cooperation and prospective reuse by agents such as human researchers, machines and hybrid collectives. Examples of communicational data are meta-data, linkages and identifiers, which have been discussed in this chapter previously.

Taken together, the four categories are the essential ingredients populating the "infra-worlds of knowledge" (Rheinberger 2011: 346) of fields such as computational and data-science. Given the relationality and interpretative flexibility of data, these categories are not exhaustive and their boundaries are not solid. They are not substantive but relational and dependent on their situational context. 'Datasets' (relational assemblages of data) in the age of open science are typically structured in a way to enable their manipulation in several categories.

Our GDP dataset, for example, is evidence for a probable spatiotemporal distribution of GDP values. It is also infrastructural data for prospective calculations and involves mobilizing data, such as meta-data, linkages and identifiers.

To assess the belonging and aptness of data in those categories is one of the essential skills that differentiate professional 'data-scientists' from other actors working with digital data (e.g. computational scientists, computer scientists, statisticians). This skill will be of increasing importance in a scientific environment subject to comprehensive 'datafication' to ensure effective and sound scientific practices. It will also require new technological devices that can help to assess data quality and make its characteristics accountable.[104]

[104] An example of such a device is OpenRefine, a "tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data." Self-description from http://openrefine.org/, retrieved on April 2, 2019.