Data curation-research: Practices of data standardization and exploration in a precision medicine database

Niccolò Tempini, University of Exeter and Alan Turing Institute, n.tempini@exeter.ac.uk

This is the accepted author manuscript of the article published in New Genetics and Society. Please cite as follows: Tempini, N. (2020). Data curation-research: Practices of data standardization and exploration in a precision medicine database, *New Genetics and Society* (in press). DOI: 10.1080/14636778.2020.1853513

Abstract

141 words

Key to precision medicine is the development of expert database projects that gather data, integrate them in the pre-existing database, and publish the product of their processing for others to make use of. Increasingly, it is required that data infrastructure managers and curators pursue and lead research projects on the data so as to learn about new ways data could be used or information that could be potentially generated from them. I call these efforts 'data curation-research' and use the case study of COSMIC, the Catalogue of Somatic Mutations in Cancer, to analyze the contextual factors shaping the science of data curation-research. I build on March's organizational learning categories of exploitation and exploration to place these factors within a theory of organizational change and innovation, and contribute to a richer picture of the social drivers of cancer genomics.

Keywords

Cancer genomics; organizational learning; big data; precision medicine; data infrastructure; data curation

1. Introduction: making haystacks to find needles

Precision medicine research aims at leveraging the flexibility that digital technology allows in storing, organizing, configuring and processing increasingly complex assemblages of data and data structures. One of its distinctive promises is to render descriptions of the world (entities, processes, procedures, therapeutic strategies) more easily manipulable, to be transferred across social contexts and conceptual scales in ways that were not possible before. In genomics, it is increasingly affordable to sequence the entire genome of a healthy or a tumor cell. Sequencing can be used to investigate if specific features of its genetic makeup are related to observed abnormalities or to suggest specific therapeutic strategies (Nelson, Keating, and Cambrosio 2013; Bourret and Cambrosio 2019). In other fields, digital systems can allow patients to log comprehensive accounts of their symptoms and experience at very specific levels of description, and researchers to aggregate disparate sets of national statistics, health care and social data (Tempini and Leonelli 2018). With data so comprehensive it could be possible, in principle, to discover razor-sharp causal relations that can be exploited or targeted through new therapeutic solutions or prevention strategies. Key to enabling these developments is the development and availability of expert databases that continually gather research data made available through the Internet by research teams, public and private institutions, and various other social actors, integrate them in a 'staging' database through standardized procedures, and publish in turn the product of their processing for other expert data users to make use of (Cambrosio et al. 2020; Leonelli 2016; Tempini and Leonelli under review).

Behind the veneer, there are important choices to be made about what gets in the database, what is left out, what is representative or relevant for the work of a research community, and who gets to be acknowledged for their work (Tempini 2015; Leonelli 2016, Prainsack 2017).

Database managers and curators are most often than not those in the position to make these choices of ‘data intermediation’. Data processing needs to be responsive to database users’ own data practices and the multiple ways in which they value data. This is, of course, easier said than done. The diversity of the user base can make value creation through data processing a particularly complex endeavour – different users put forth different value questions to database managers (see Tempini 2017). As Green, Carusi and Hoeyer argue (2019) argue, the idea of developing ‘finer-grained taxonomies’ (as prefigured in the precision medicine imaginary) is never purely an epistemic issue. Rather, it exemplifies the intersections between “epistemic, organizational, regulatory, and political issues” (2019:7). When it comes down to making data re-usable, data classifications can be negotiated in both directions. Sometimes, coarser is better (Tempini 2015).

Leonelli (2016) argues that cutting-edge big data research requires for data to be ‘labelled’ and ‘packaged’ with various kinds of metadata as a precondition for their successful reuse in new situations of research. Data curators, who perform the labelling and packaging, aim to provide information about, among others, the provenance of the data, the original context of measurement and use, and the object of observation. The intention is to facilitate the re-use of the data, but because of the performative normativity implied by activities of classification, categorization and contextualization, data management and curation also shape or frame the way in which the data are thought of and used (Leonelli 2012) – curation is not epistemically neutral. While data curation does not fall in what traditionally has been considered research, it should be given more attention in the study of epistemic practices.

In this paper I want to build on this argument and extend it to account for recent developments in the way databases reach maturity and find a path towards their long-term sustainability. I will argue that to make research with health big data possible it is increasingly required that data infrastructure managers and curators *pursue and lead research projects* on the data so as to learn about new ways they could be used and information that could be generated from them. This translation re-organizes identity and structure of datasets through various operations of integration, aggregation and calculation (Lee et al. 2019; Tempini 2020). Lee and colleagues characterize the "translation not transmission" (2019:9) operations that algorithms perform as "folding": manipulating relations between data points and consequently re-organizing the data in new representational formats, so that such relations are elicited and made amenable in the situation of inquiry. However, while 'folding' attends to the re-organization of data relations, datasets are objects defined by several socio-technical relations (Tempini 2020). As source datasets are disassembled and re-assembled to produce new derivative datasets, other epistemic practices and organizational processes need to fall into place for the data to have consequence in the situation at hand. If the data do not speak for themselves, we may say, nor do algorithms.

Accordingly, the research projects this paper is interested in see curators ask themselves questions about the potential semantic value of the data, and the emerging trends in scientific research that should be as much as possible anticipated and organized for in the struggle to keep their work relevant and ensure their infrastructure project a sustainable future. Pressures on data curation projects (including changes in funding and industry structure) put curators in the need to lead and show the way for new paths of data re-use. Databases are complex epistemic objects that are not easily understood, interpreted and re-used, and curatorial teams have some of the best expertise available in navigating them (Tempini and Leonelli 2018).

I suggest calling these efforts *data curation-research*. Even if they rely heavily on operations of classification, categorization and aggregation, the practices of data curation-research are set apart from labelling and packaging processes, which sit in an intermediary space in respect to research processes traditionally defined. Such curation activities shape the epistemic value and fate of scientific data, while at the same time are limited to an 'infrastructural' role that is more indirect and less invested in specific research endeavors or applications (Bowker 2000; Leonelli 2016). Data curation-research is instead more akin to the imagining of a new knowledge infrastructure rather than to the management of an existing one.

Studying the rise in cancer genomics of a new kind of infrastructural resource, the knowledgebase, which builds on top of databases and is increasingly key to enable data re-use, Cambrosio and colleagues (2020) observe how a renewed emphasis on the further generation of new interpretations makes a strict focus on labelling and packaging practices a difficult fit. They point out instead to the several layers of infrastructures and data integration practices that are required to make cancer genomic data usable in the clinic. Similar to the way in which knowledgebase projects are meant to unblock the 'interpretation bottleneck' in the clinic, data curation-research is centered around exploring new data re-use possibilities. It builds upon the work of data labelling and packaging, but includes a wider set of practices that are at the same time aimed at 1) addressing a research question or task, and 2) trace new 'paths' delineating how data might be used in new ways, releasing a legacy of technologies, methods, derived data, and practices that will be more easily employed by subsequent internal and external users of the data. Key is the development of new data products that can eventually be taken up in a research field as a standard resource. Data curation-research, I will show, is an endeavor that many mature database projects seeking long-term sustainability are likely to pursue. These projects need to perform a double act, on the one hand, maintaining continuity and standardization of existing data sources and associated processes, and on the other hand, engaging in explorative processes that potentially depart from established practices.

In this paper, I present an example of data curation-research in the case study of cancer genomics database COSMIC (see also Cambrosio et al. 2020). The study is based on qualitative data collected through interviews, participant observation, and analysis of secondary evidence,[1] and employs the conceptualization of the 'data journey' (Leonelli, 2016) as a methodological strategy to reconstruct the web of interdependent socio-technical practices enabling the sharing and re-use of data. The analytical lens of the data journey can be a fruitful angle to make sense of the consequences of data practices (see Leonelli and Tempini 2020). It assumes that data move through many steps across different social settings before they reach the end use. Through their movement, many interactions take place that will shape the meaning, epistemic value, and future uses of the data. The data journeys angle

[1] Fieldwork was carried out between 2015 and 2017. The twenty-one interviews were semi-structured and followed a questionnaire that was customized for each interviewee after in-depth preparation. They were accompanied by participant observation on site over two visits equivalent to six days (full-time). This included attending an away day where members and external collaborators presented their work on COSMIC and reflected about the state of the project. Further informal discussions were carried out in informal socialization and in the occasion of a COSMIC away day I participated to. The research is also informed by secondary evidence gathered from the COSMIC website and the vast cancer genomics literature related to COSMIC. Interview transcripts are available for more detail, open access (see 'Data Availability' section).

requires the social analyst to try and reconstruct the histories of data, and account for the ways they matter.

To understand what factors can shape the development of data-curation research in response to interaction with collaborating organisations and other factors of COSMIC's operational environment, my reflection builds on March's seminal work on adaptation and organizational learning (March 1991), which sees organizational survival as dependent on maintaining an appropriate balance between the opposing activities of *exploitation* and *exploration*.[2] The organizational learning problem that characterizes COSMIC's case, it should be clear, is *how to keep the database valuable for third party users? What data processing practices for added-value database development should the team pursue?* March defines exploitation processes those which refine and extend *"existing competencies, technologies and paradigms"* (85) and are concerned with *"refinement, choice, production, efficiency, selection, implementation, execution"* (71). Exploration processes instead pursue new possibilities and are concerned with *"search, variation, risk taking, experimentation, play, flexibility, discovery, innovation"* (71). Exploration is 'generative' (Aaltonen and Kallinikos 2013) and opens new directions of development, but is also redundant and inefficient. Exploitation is dependable and focused on the preservation and improvement of the value generation of existing processes, but is also passive and repetitive.

[2] For a relevant antecedent (and inspiration) see Aaltonen and Kallinikos' case study of coordination in Wikipedia, which also employs March's taxonomy of organizational learning (2013). However, the nature and definition of Wikipedia's 'knowledge making' and organizational learning that the authors are working with is not always clear for the purposes of this paper, and their study does not share the specific focus on data practices that I work with.

Most organizations are supposed to operate both learning modes at the same time, albeit to varying degrees, as they are involved in multiple and heterogeneous endeavors. Finding an appropriate balance is key to *"system survival and prosperity"* (March 1991: 71), but it is difficult to know just what balance a specific organization needs to strike. The outcome of organizational learning is highly contextual and dependent on a number of conditions including 1) the degree of knowledge heterogeneity and learning convergence between the organization's members, and 2) the degree and kind of environmental change and shifts occurring in the institutional ecology the organization operates in. In a fast-changing reality such as that COSMIC operates in, performance may be improved by strategically managing team turnover, and controlling learning convergence between team members. By ensuring there is turnover and hires are heterogeneous to the existing talent pool, exploration is stimulated as it becomes easier to embark on experimental projects, and the organization better prevents the risk of falling out of sync with changes shifting the field and, consequently, user needs and priorities. At the same time, the level of convergence between different organizational endeavors should be controlled. If all members share in the exploration of new alternatives, risky experiments could jeopardize the stable rewards that older, standardized ways of working have secured. If convergence is appropriately limited, experimentation can be more easily 'incubated'. Managers should pay attention to the way incentives are distributed, if some members are allowed to embark in exciting experimentation and to learn faster, while others need to ensure consistency of their work process, and learn slower. For these reasons, it can be useful to keep side projects separate from core production processes: *"multiple, independent projects may have an advantage over a single, coordinated effort [...] variability can compensate for the reduced mean in a competition for primacy"* (March 1991: 84).

In the next section, I present the case study of COSMIC, giving particular emphasis to shifts in the organizational processes, structures, and information product innovations, as they happened in response to changes in the environment in which COSMIC operates: trends in the cancer genomics field, and changes in funding source and composition of the team, among others. In the third, discussion section, I reflect on the relevance of organizational factors for the debate on the epistemology of data practices and in particular of data curation-research, drawing on James March's seminal work to explore the intersection between organization studies and data studies.

2. COSMIC

The Catalogue of Somatic Mutations in Cancer (COSMIC)[3] was born from a cancer genetics list, in the form of a spreadsheet curated by some of the Principal Investigators of the Sanger Institute's Cancer Genome Project (CGP), based at the Wellcome Genome Campus in Hinxton, a few miles outside Cambridge (UK). The PIs needed a resource about the genes that were known to be implicated in cancer that could be shared internally. The spreadsheet recorded and formalized their first-hand research experience and knowledge of the literature and counted several dozen entries. In 2004 this resource was elevated to a full-fledged project, centered around a website that would be regularly updated in order to keep up with the discoveries that were accumulated at increasing pace in the field of cancer genetics. The launch website published the data aggregated from 1483 papers about 4 genes, supplied a

[3] Somatic mutations are acquired during one's life and are to be found only in some cells, as opposed to *germline* mutations, which are inherited and found in all of the body's cells.

new tumor classification system in the absence of a pre-existing standard, and positioned itself against other cancer databases as the only database that was not restricted to one gene or location (Bamford et al. 2004). COSMIC was born with the explicit aim of enabling the circulation of genetic evidence from cancer research amongst the research community. Over time, the database was adopted by very different kinds of users, which today range from targeted molecule pharma developers, diagnostics researchers, to clinical users taking decisions on patients' prognosis and treatment course (Tempini and Leonelli under review).

The field was quickly shifting from genetics to a genomics angle, as the first large next-generation sequencing projects pioneered new approaches to the study of cancer. Cancer genetics data were shared through individual publications and that reported on specific relationships between a limited number of mutated genes and cancer events and outcomes. Next-gen approaches aimed at sequencing whole genomes of cancer tissue samples. This goal would not to be reached all at once, but stepwise – initially sequencing was mostly limited to exome or other genome subsets – but it represented an important methodological shift nonetheless, not least because new studies aimed to produce statistically relevant evidence, which introduced new sampling requirements. As the challenges in operationalizing and funding the new approaches were big, progress was led by large international consortia. International Cancer Genome Consortium (ICGC) was launched in 2008 to sequence the genomes of 500 cancer samples for each of 50 cancer types. The Cancer Genome Atlas (TCGA) started in 2005, to sequence dozens of cancer types with a similar amount of samples. These studies started to share vast 'raw' datasets recording sequencing information for each of the patients involved in the study. Since the first few years, the project needed to adapt to the trend and prepare to include both cancer genetics and cancer genomics data sources. By 2005 COSMIC started to process data sources from two different streams

(Forbes et al. 2006), literature data about 28 manually curated genes, and mutations for other 518 genes extracted from screens data published by the CGP. By 2011, genomes from TCGA were being imported (Forbes et al. 2011), and by 2015 ICGC genomes were also added (Forbes et al. 2015). By then, genomes from consortia constituted about half of those curated into COSMIC. The other half were genomes sourced from individual publications – next-generation sequencing had become feasible for many more groups and projects.

[FIGURE 1 HERE]

Today (see Fig. 1) COSMIC is a fundamental resource in cancer genomics, and is used on a daily basis in research hospitals and cancer centers, and diagnostics and pharmaceutical companies, where it comes into play as one of the key sources to interpret data from sequenced samples: the crux of the genomic enterprise has not been how to obtain genetic sequences, but understanding what they mean and what to do about them (Nelson, Keating, and Cambrosio 2013; Cambrosio et al. 2020; Timmermans et al. 2017). The COSMIC database is used as a reference point to juxtapose locally-generated sample sequences, so as to highlight links with observed cases: for instance, one can observe that a particular mutation, reported in the sample, is linked with a high number of cases of a specific type of cancer affecting a particular organ. COSMIC data can be browsed through a web analytics interface that allows to investigate the data at the level of cancer tissue type, the specific gene and mutation, and through whole genome browsing. These multiple visualizations support several para-digmatic ways to explore the same data (see Manovich 2001), for instance traditional tissue-centred heuristics (distribution of mutations in cancer of the liver vs. of the lungs) vs. gene-centric investigations (distribution of mutation and cancer types in association with a gene vs. another).

Consider mutated KRAS gene, one of those most implicated in cancer. The COSMIC Gene Analysis page for KRAS (https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=KRAS) starts with a Gene View section providing a series of histograms that visualise the distribution of observed mutations, divided by mutation type, across gene locations; and corresponding gene expression data. Both mutation and gene expression visualisations can be clicked through to pages where data tables list information for all samples where the event has been observed. Each sample from the list can be probed further for reference information. Mutation pages are also provided with tissue distribution counts. Clicking on tissue distribution counts leads to a page from the Cancer Browser that summarises main mutations involved in that particular type of cancer, and other pieces of information. The Overview section provides summary information, including sample counts, gene coordinates, protein product counts and 3D visualisations to be further explored through COSMIC 3D (more follows), data about drug sensitivity. A link for KRAS as a cancer 'Hallmark Gene' leads to a dedicated page publishing a summary of the functional features of mutated KRAS, the key categories of its behaviour (for instance, 'angiogenesis', 'cell replicative immortality', 'escaping programmed cell death', among others), and links to key evidence. The External Links section provides links to information in other repositories, databases and genome browsers materials. A Tissue Distribution section provides descriptive statistics of the tissues affected by KRAS-mutated cancers across the observed samples. Similarly, the Mutation Distribution section provides counts and pie chart visualisations of the types of mutations observed. A References section lists publications that COSMIC staff associate with mutated KRAS (including links to publication pages on Pubmed and COSMIC). If the publication is categorised as 'Curated', its COSMIC page will provide information about mutation locations, genes, variants and samples that have been extracted it by staff.

The database can also be entirely downloaded for the most granular analyses by researchers using custom software setups. Eliciting these pieces of information is a key step in composing the extremely complex picture of informational cues that experts (or boards of experts called Molecular Tumour Boards – see (Bourret and Cambrosio 2019) need to evaluate and decide upon prognosis and treatment course.

New information needs, generated by the shift from genetics to genomics, shaped the project's operating epistemological, organizational and logistical assumptions. While the raw capacity to generate sequences 'without interpretation' has multiplied thanks to disruptive methodological and technological breakthroughs, the field has struggled to keep the pace in the formulation of hypotheses and explanations of the causative processes of cancer. Through the dominance of the oncogenetic theory of cancer causation (Keating and Cambrosio 2011), cancer has become the default example of a disease that is almost always not monogenic[4] and yet is strongly related to complex combinations of genetic mutations, gene networks and activation pathways. While a minority of mutations are well understood to either cause or predict cancer, many observed mutations that are associated to at least some cancer cases are putatively implicated but, as COSMIC's director Simon Forbes observes, *"no one quite knows how or why".* The picture has quickly become extremely complex. New high-throughput genomics studies became possible that generate data in a 'broad sweep' and leave the interpretation of the sequences as a second step. How to piece together the gene network 'puzzle' is then the outstanding problem (Cambrosio et al. 2020), but it is now widely accepted that even a focus on gene networks is too reductive if it neglects to account for

[4] Refers to a state of things where mutations on a single gene cause the disease.

cellular processes of expression and gene regulation. Networks must include various kinds of molecular actors (including industrially made ones – see Vignola Gagne et al. 2017) if they are to shed light on cancer evolution.

And yet, genomic data are disseminated in highly diverse and non-standardized ways (for instance they can be shared inside an article, as publication appendix, as datasets, etc.). Moreover, the data shared in cancer genomics research has continued to grow not only in size but also in diversity, with new kinds of data added over time as studies operationalized increasingly complex hypotheses of cancer causation – to mutation data, COSMIC has included, among others, methylation data on gene expression and as well as drug resistance reports. Rob McEwen, pharmaceutical researcher from Astra Zeneca, provides one use case example revealing how deeply embedded the database is in the contemporary infrastructure of cancer genomics: *"In order to get a sense of the prevalence of a mutation in a cancer type you need quite a lot of samples, so it's not something we do in house. [...] So it's finding those pieces of information from which you can then make those sort of inferences. That's where COSMIC is useful, because it is like a starting point, a one stop shop to dive in and look for that information. [...] So you can see, "Oh, right. This mutation's also in COSMIC, so it's obviously been seen in the public domain. It's not completely unique." [...] Then we go in and review that and start to drill down into which ones might be relevant based upon the background, so the particular hypothesis or the situation in which the tumour was taken."* To collect, standardize, aggregate and re-publish research data is a key fundamental step for organizing knowledge and sharing it. Operating in a fast changing environment, COSMIC found itself, time and again, pulled between the need to keep data curation practices consistent and reliable (what I will call: data practice standardization), and the need to create new data curation practices that respond to shifting trends in the field (data practice innovation).

2.1 Standardized data practices

COSMIC negotiates its intake of the gargantuan amounts of data made available globally every day through two organizational processes that the team has tried to standardize as much as possible. The first standardized process is expert curation of cancer genetics literature. Through this process the team exhaustively curates the literature on a list of key cancer genes. Post-doctoral level curators with a background in biology extract data from the papers they read, and they input them through a web interface in the database's working copy. This process is suited to aggregate smaller quantities of high quality, deep data on observed mutations of specific genes of interest, and the cancer events and signs that were observed in association to them. It has been a key proposition since the launch of COSMIC in 2004.

The genes that are so curated by COSMIC are a subset of all those that might be implicated in cancer, and no statistical principles of sampling can be applied in practice, in curating the deluge of cancer genomics data. The list reflects the team's understanding of research trends and discoveries as it tries to cover the genes with the strongest evidence in support of their putative role in cancer causation. Once a gene is selected for inclusion in the list of manually curated genes, the team waits to release data about it until enough of the literature has been curated to have the perception that the main events and mutations associated with the gene have been well captured in the data that will be published. The new gene curation is then announced in the occasion of a new COSMIC release and after then, the curation is continuously maintained – curators will keep integrating new data as it is published and depending on its degree of novelty.

To optimize both curation consistency and productivity, the list is split among the team members so that each curator specializes on the curation of the literature of a fixed subset of genes. New hires start by reviewing the literature of one or a few genes until the curation is published. Over time, they can pick up new curation responsibilities of other genes' literature, while they maintain the curation of previously curated genes. Experienced curators 'keep the pulse' of research on dozens of genes at once. This level of manual curation of the literature and the data extraction it involves is extremely labor intensive. An economy of scale of sorts is enabled by repeating curation of same gene over time. At the same time, an economy of scope of sorts is also enabled as the choice of new genes to curate tends to fall on genes that are neighbor to the ones that a curator is already curating: neighbors are genes that tend to be studied, or show up mutated, when the first gene is. The standardized operations and strategies of the manual literature curation process are the result of many years spent integrating high quality, deep genic data in a fast changing environment, where research trends and the underpinning understandings of the processes driving cancer keep shifting (S. A. Forbes et al. 2015).

The second standardized process negotiates the intake of a different kind of data source, broad genomic data. It involves semi-automated importation of whole-genome and gene-panel datasets. Post-doctoral level curators with programming skills download genomes from the repositories of big panel studies such as The Cancer Genome Atlas, and with the help of software tools integrate these datasets in the COSMIC database. This process is suited to intake broad, comprehensive data on many variables and patient specimen at the same time. The addition of genome data integration was a response to the increasing availability of full datasets from high-throughput studies, and the resulting changes in expectations from end users – with the shift to genomics, research became increasingly focused on the generation of most comprehensive gene panels as a first and routinized step before the execution of investigations motivated by specific working hypotheses.

New data are gradually integrated in the database over a course of three months, until they are distributed as a new database release. The two standard processes of data curation have been refined over a relatively long time, and curation is made consistent at the team level through practices of team review and supervision of the new hires by the most experienced members. There is strong emphasis for curators to follow the organizational standard on how to prioritize curation between different data sources, how to judge their quality, and how to select the genes to be curated next. The team of curators is also rather stable in composition with low turnover. Curators are employed on a long-term and enjoy workplace flexibility, with several working from home (but for key meeting days). This scientific career offers a trade-off that can be attractive, and several members have held the job for many years.

COSMIC has become a key component of the global cancer genomics infrastructure in large part thanks to the stability and consistency of the two standardized curation processes that build the project's core information product, the database of cancer somatic mutation data. It is important to place standard organizational processes in the larger context of its industrial and financial environment, and crucial to COSMIC's achievement has been the low turnover of the curatorial team. This was made possible by the charitable funding of the Wellcome Trust, which has provided a stable source of income to employ curators long-term and develop the project's core product, and the freedom to focus on product refinement with relatively little concern for other organizational drivers such as market position and strategy.

2.2 Innovative data practices

After more than ten years of steady growth, the Trust capped its charitable funding in the late 2010s. COSMIC management was encouraged by the charity to explore industry partnerships as a way to fund the further growth that the explosive development of the cancer genomics field made necessary, and introduced a tiered data licensing model.[5] The first industrial collaborations have allowed the team to double in the span of a couple of years, providing project funding for the fixed-term hire of a number of researchers tasked with exploring new ways that COSMIC data can be put to valuable use according to the interests and competencies of the industrial partner. For the partner, the possibility to leverage the intimate knowledge of the COSMIC database held by in-house researchers can be an attractive proposition. Uncertainties about the context and provenance of data shared through web databases often undermines their re-use, and collaborating experts with first-hand experience remains one valued solution (Tempini and Leonelli 2018) for companies who want to invest in the database as a key research resource and need to secure its alignment with their organizational processes (see the next subsection, 2.2.1, for the example of a pharmaceutical developer sponsoring a project to use COSMIC mutation data in the exploration of the structures of mutation protein products). As a result of these developments, the organizational structure of COSMIC has grown more complex, with a multiplication of roles and commitments. At the end of the 2000s staff were only a handful of people between management, curators, and software developers. When in the 2010s the project started to deal with the heavier curational demands of genomic data, charitable funding grew. By the time the business model was opened to alternative funding streams, in the mid 2010s, charity

[5] The data are now available for free for not-for-profit use, while for-profit users pay an access fee, while before all data were released free regardless.

money employed nine staff. Now, a permanent team of curators focused on the standardized development of the core product is flanked by a high-turnover resource pushing the boundaries of COSMIC data use in new directions. By the time, in the late 2010s, that several private sector collaborations were on, the team had grown to between 20-30s. It included management and project managers; several software developers, database specialists, and core data curators; and a handful of postdoctoral researchers focused on specific research projects.

The fast development of the cancer genomics field forced the COSMIC team to interrogate themselves as the data resourcing needs of its users are changing. The fear is that the project loses the standard position it conquered, and with it, further funding. While on the whole deemed an inevitable development for COSMIC, industry partnerships have also carried risk. On the one hand, they provide the team with a direct source of intelligence about user needs and research trends. COSMIC management stressed the difficulty, for a publicly available database project, of obtaining information about data's end uses. On the other hand, the team has been careful about marrying the project to the specific needs and priorities of its partners at the expense of the universality of its product and image. As Forbes highlights in retrospection, the partnership model has been considered very successful in enabling COSMIC to maintain relevance and renew its data product offering in the face of a continuously changing field: *"We started partnerships with other companies on discreet projects analyzing across that data, and that's been very interesting because it takes us away from our core capability and allows us to decorate the core capability with knowledge derived from these additional analytic collaborations. [...] Then we knew there's value behind that because it's already impacting the research of that company, and if it's impacting the research of that company, other people will want to know that sort of thing."*

Trying to avoid finding itself between a rock and a hard place, it has become clear that key for the project's long-term sustainability are 1) using the data first hand, and 2) *recursively* sharing the *new information products* that have been developed in the process.

2.2.1 COSMIC 3D

One example of these forays in data exploration and curatorial research has been COSMIC 3D, a new web-based data visualization feature that shows, through 3D visuals, the effects of a mutation on the consequent protein product: does the mutation 'fall' on the binding site of the protein, thereby becoming a potential target for pharmacogenetic compounds? COSMIC 3D was developed in collaboration with Astex Pharmaceuticals, a drug developer specialized in the design and development of targeted molecules, thus combining different expertise: Astex contributed a "structural biology perspective" on oncology, COSMIC a genetic one.. COSMIC 3D offers its users a way to understand functional implications of mutation by examining differences between the protein products of wild type genes as opposed to those of a specific mutation, and evaluate the relative opportunity for therapeutic intervention through targeted small molecules. To develop it, COSMIC hired a postdoctoral researcher, Harry Jubb, who was educated nearby in Cambridge (biochemistry) and specializes in structural biology bioinformatics approaches for drug discovery. He had previous research experience linking COSMIC data to investigate the structural biology of drug resistance in collaboration with key Astex researchers, and was keen to experiment with new computational infrastructure (e.g. MongoDB, a noSQL database popular for large scale distributed systems). In COSMIC 3D he was tasked to analyze the cancer mutation data available in COSMIC, link them to protein data available from another public repository, Uniprot, and develop the visualization feature according to the way pharmaceutical developers may use it. In their own reporting, the curator-researchers emphasized the potential for COSMIC 3D to suggest, explore and provisionally validate novel hypotheses (Jubb et al. 2018).

A researcher wanting to pursue a particular line of reasoning can explore this kind of '3D jigsaw' in many ways through the refinement of sophisticated queries that select only particular data and visualize the results accordingly. The underlying rationale, an interviewee explained, is to *"understand what's driving cancer by seeing how it changes the structure and the function of the protein that the cancer genes make"*, and for this reason COSMIC 3D is seen as a technological resource part of a broader research endeavor. By linking together datasets that have not been linked before, COSMIC researchers are able to pre-calculate the combinatorial space of the possible spatial arrangements implicated by mutations. Ranked by an algorithm, results can be flagged for further examination in the drug design context. Here the suitability of the flagged result for further drug development actions can be examined in relation to the portfolio of available molecular tools, with the hope to determine that *"there's potential for [this mutation] to be targeted because it's next to a binding site"*. Used not only for in-house research, this technology for hypothesis generation "*is very powerful because it's very visual, you can explore, it's being used inside Astex and we know there are lots of people over the world using it from the analytics that we have."*

2.2.2 Cancer Gene Census 2018

A similar initiative is COSMIC's 2018 version of the Cancer Gene Census and its features of 'Cancer Hallmarks' review, which involves the publication of a summary review of the carcinogenetic roles and implications of mutated genes (Sondka et al. 2018). The feature builds on a pre-existing resource – the Cancer Gene Census (CGC), a list of the genes most implicated in cancer. It was launched in 2004 by researchers involved in the Cancer Genome

Project of the Sanger Institute (Futreal et al. 2004) and featured 291 genes showing mutations, in cancer samples, with the most likely relevant functional consequences. The CGC has since become a *de-facto* classificatory standard in cancer genetics research, medical reporting and drug development (Sondka et al. 2018), and the 2004 paper has been cited more than 2800 times. Since COSMIC's launch in 2004, the CGC has been a key organizational reference point for COSMIC curators, who maintained it up to date and used it to decide what other genes' data to take up full curation of, as they extended the selection of genes that COSMIC ought to include. The CGC list was, however, very limited. It included no explanation as to the reasons for the inclusion of each gene. The 2018 CGC overhaul aimed to right that. It required more than 2 years of project work and a post-doctoral researcher, Zbyslaw Sondka, tasked to lead it in collaboration with colleagues including core COSMIC curators. He has a background in biochemistry and first-hand lab experience in cancer genetics research, which is key for assessing the details of genetics studies, but had also developed experience in the bioinformatics of next generation sequencing. His two-fold background provided him with relevant experience to assess the different kinds of evidence that the CGC reviews (more below). The aim of the new CGC was to completely re-generate a state of the art review on the causes of cancer and formulate new statements about them and their general functional features.[6] As Sondka highlighted*"curators are focused on different issues from this. So, the set of literature covered by COSMIC doesn't have to completely overlap."* The new CGC required new more functional information about mutated genes. It needed to include genes mutated in the germline (as opposed to somatic), and to address how to curate and categorize gene fusions.

[6] Notably, none of the researchers involved had been part of the team that authored the 2004 CGC paper, which was developed in a different group at Sanger.

To some extent, the CGC 2018 was a sort of systematic review. It was not anything like the standardized curation that COSMIC has pursued for 15 years. First, it fully re-assessed the evidence available about all the genes already included in the CGC, adding and removing genes from the list according to the most up to date information. It included validation criteria such as evidence review by two independent COSMIC researchers, and the requirement of the existence of multiple sources of independent evidence to warrant inclusion in the list. Second, it introduced the requirement of two different types of evidence for the qualification of a gene as unequivocally implicated in cancer: 1) the existence of robust mechanistic explanations; and 2) the availability of substantial data showing anomalous mutation patterns in the COSMIC database. Priority was given to the construction of reviews from, as an interviewee put it, *"collective evidence from different sources, possibly from many papers, possibly from many different experimental techniques."* Researchers thus explored the mutation data that had been painstakingly accumulated in COSMIC from tens of thousands of papers as an evidence base to reach conclusions about the experimental support to a hypothesis of gene implication. If only one of the two types of evidence is available, the gene is included in a new 'second tier' list of the census, where genes with compelling but ultimately inconclusive evidence of causal implication are included.

After reviewing the evidence, COSMIC researchers compiled a summary review page (for more than half of Tier 1 genes) that provides a gene 'profile', qualifying the gene against established categories of cancer gene action (e.g., invasion and metastasis, suppression of growth, proliferative signaling, angiogenesis, etc.). These classifications utilize the well-established characterization of 'cancer hallmarks' (Hanahan and Weinberg 2000), but also expand on it, categorizing genes for other biological processes such as *"cell division,* *differentiation, global regulation of gene expression, senescence and impact of mutation on gene or protein function"* (Sondka et al. 2018:699). Through this profile, a comprehensive review of all the roles a mutated gene can take in cancer is compiled, which, according to the authors, *"can highlight unexpected and targetable aspects of [the gene's] activity"* (699). They elaborate that *"cancer arises through complex multidirectional modulation of cellular processes associated with hallmarks of cancer rather than through simple promotion of all hallmarks"* (700).

The CGC 2018 review published in *Nature Reviews Cancer* produced new knowledge and highlighted new avenues for the development of precision oncology. First, the introduction of a second tier has established a category for an accelerated identification of genes *"with strong indications of a role in cancer but with less strong mechanistic or functional evidence"*; of genes with strong mechanistic evidence but unclear mutation patterns; and of genes regulated only epigenetically (698). Second, it discussed *"some of the striking observations that are possible only via such a broad, deeply descriptive curation effort as the CGC"* (698).[7] The authors argue that it is only through "large-scale deep curation [...] that the sheer extent of [context-dependency in the roles mutated genes take] can be appreciated" (700).

[7] Among the highlights are the breakdown of the traditionally sharp distinction between 'oncogene' and 'tumor-suppressor' gene (gene roles become more fluid and contextual to a number of other variables – some genes can act as tumor promoting under some conditions and as suppressor in others). At the same time, different mutations on the same gene can have distinct impacts on oncogenesis; different tissues and tumor stages can give the same mutated gene completely different roles; genes can participate in cancer-driving networks through global epigenetic patterns.

[TABLE 1 HERE]

4. Discussion: exploitation and exploration in data science

The case of COSMIC illustrates the long-term questions facing a database project in a fast-moving field such as cancer genomics, which has been spawning an increasingly diversified research industry intent on developing targeted drugs, diagnostic methods, health care products and services. The value of data is never settled, as their semantics change each time records are aggregated, computed and re-organized in response to shifts in the organization's understanding of the field and the subject matter of cancer research. The case study shows how the COSMIC organization has itself changed over time, re-arranging its data curation work processes to be able to, on the one hand, maintain and update its standard, backward-compatible, *core data product* the COSMIC database of somatic mutations, while on the other hand, to perform explorative forays (data curation-research) that can lead to *new data products,* the development of new data work processes and the production of new knowledge (see Fig. 2, Table 1). Beyond merely re-stating the fact that big data practices are not happening in vacuum (but rather, the meanings and uses that data are brought to bear are dependent on the context a project operates in), this case study gives us the opportunity to reflect on how this database project has matured and the organizational conditions that have allowed it to adapt to change while maintaining its standard making position.

[FIGURE 2 HERE]

March's theory of adaptation and organizational learning can help us explain key developments in the case study and the relationship between core and new data products.

COSMIC's trajectory, since the early days of the genomics shift to the present, has been shaped by a number of external factors that have forced management to think strategy carefully, including changes in funding streams (capping of charitable funding forcing the search for alternative sources to sustain growth) and increasing complexity and dynamism of the maturing field of cancer genomics. Management has recognized the need to generate new knowledge about the environment the organization operates in and launched a number of explorative projects, which have been incubated (developed on the side from the core product development activities) and gather input externally through engagement with collaborators and employing new talent (who are given a circumscribed opportunity to embark in higher risk/reward projects than routine curation can offer to permanently employed members). Knowledge about COSMIC's environment is not limited to knowledge about other organizations (customers and competitors) and their projects, but also 'knowledge about knowledge', of cancer and its trends, which is key to be able to predict shifts in the field and maintain COSMIC data's standing as a standard maker. For these reasons, explorative projects in COSMIC take the form of *data curation-research,* and are characterized by the re-use and re-combination of existing COSMIC data sources with one another (CGC 2018 – for the most part) or with new external sources (COSMIC 3D), for the purpose of addressing a set of research questions and discovering new information that can be shared as *derivative datasets* (see Tempini 2020). These new datasets can be used by external researchers to explore and ground their evidential claims, and can become the source material for new data explorations.

[FIGURE 3 HERE]

In turn, the case study can help us to explain what use the organizational learning categories of exploitation and exploration can have in the social study of data science and work. It allows us to link organizational learning to scientific discovery, and understand the ways in which specific organizational factors sustain the database enterprise over the long run. Standard makers such as COSMIC are in a difficult position. On the one hand, continuity, comparability, backward-compatibility are key for the database to become installed base in the cancer genomics infrastructure. This provides a strong driver for the organization to adapt its learning strategy to exploitation (see Fig. 3). The curation process will need standardizing. The economies of scale and of scope in the curation of gene literatures (see section 2.1), the continuity of the curation team membership, and phased training of new hires, will increase process efficiency. The data that will be produced will be most reliable considering the level of curation refinement and specialization. On the other hand, to be standard infrastructure for a field of scientific research and industrial applications as fast changing as cancer genomics requires the project to try and anticipate not only changing inclinations, priorities, and competitive strategies of the many users, but also knowledge gaps and discovery trends in cancer research itself. Regardless, COSMIC needs to balance exploitation with exploration, and experiment with projects that are incubated on the side to test the potential of their research concept in practice, drawing from the available curation and research know-how as appropriate but without interfering with core curation processes. Only once mature, new learning generated by experimental projects is embedded in the rest of the organization's processes, products and infrastructure. Too fast learning from experience risks increasing uncertainty (March 1991). COSMIC projects could be risky, if they were to take the team to develop research and resources that turn out to be not of much interest for the field at large, but for the sponsoring organization.

To conclude, the lens of exploitation vs exploration allows us to identify and articulate some of the key organizational conditions to enable value-creating data curation practices, and move beyond the well-trodden observation that closely knowing the data, and the characteristics of the situation of use, is crucial for successful data curation. It allows us to interpret the contextual factors driving data curation-research activities, place them within a theory of organizational change and innovation, and provide a richer, 'holistic' picture that accounts for the intersections between epistemological and organizational issues.

In respect to the future of COSMIC, the case study suggests that data curation-research forays, and the learning cycle of data exploration and exploitation I have elaborated, are not constrained to one-off explorations. Rather, they can become the operational model of standard-making organizations in data-intensive fields. COSMIC Director Forbes sees in these projects a unique and persistent opportunity to secure the learning opportunities that keep the database evolving and relevant: *"Every time we enter into a collaboration or a partnership, it takes us further away from our comfort zone and further into much more exciting new ways to push forward the precision medicine aspects of the work that we do."*[8]

[8] Forbes observes that the organization is well placed to exploit the position it holds for the re-use and interpretation of its own data: *"If you consider the core COSMIC database as the data, we're trying to build knowledge on top of that, that assists people in interpreting that data, because there's a limited number of bioinformaticians across the world who are either interested or capable of just scrutinizing such a gigantic data source. And even then, you have to understand the data, its context, its foibles, in the broader context of bioinformatics and it just takes a while."*

Ideally, exploration-exploitation learning cycles should follow one another *recursively.* With datasets adding up and data types multiplying continuously, questions of the relevance of data will keep spawning. An open issue is of course at what conceptual scales, and with what organizational arrangements for data analysis is data curation-research most fruitful. Other initiatives are addressing similar problems of data interpretation, prioritization and meaning-making through different organizational models, as Cambrosio and colleagues show in their case study of knowledgebases (2020), which are supposed to sit on top of databases including COSMIC. CGC developer Sondka expects that regardless of the focus and direction, new data-curation projects in COSMIC will keep building from the outcomes of previous data curation-research projects. The picture *"is going to evolve even more because Harry has finished his COSMIC-3D. We are starting work on the mutation census, so Sam is doing methylation analysis, so they all are going to interact. [...] it's all going to mix up and we don't even know what's going to happen next."*

Acknowledgements

I would like to thank the Editors and members of the reviewing team for their invaluable comments and support to the development of this article. This research would have not been possible without the generous collaboration of my informants and the team at COSMIC led by Simon Forbes – thank you all. Early versions benefited from the feedback received from audiences at the 2018 SPSP conference in Ghent (special thanks to Sara Green) and the 2019 ISHPSSB conference in Oslo – thanks to Sara Green and Steve Elliott for leading the organisation of our sessions. Thanks to Hans-Jörg Rheinberger for offering comments in the latter event, and to David Teira for comments on a late manuscript draft. Last but not least I thank Alberto Cambrosio, Pascale Bourret, Henri Bergeron and Patrick Castel, and the authors participating to the workshop that they organized at the Centre for the Sociology of Organisations – Sciences Po, Paris, for providing the context that made this article possible. The first draft was typed in the reading room of the Cahersiveen village library in Co Kerry, Ireland – unlikely circumstance and lovely memory. This work was supported by ERC grant award 335925 (DATA_SCIENCE) – thanks to Sabina Leonelli – and by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

Declaration of interest

The author reports no conflict of interest.

Data availability

Interview transcripts are available as open access dataset on Zenodo:

Tempini, Niccolò. 2020. "[DATA_SCIENCE] Interviews: The Catalogue of Somatic Mutations in Cancer (COSMIC)." Zenodo. doi:10.5281/zenodo.3944466.

References

Aaltonen, Aleksi, and Jannis Kallinikos. 2013. "Coordination and Learning in Wikipedia: Revisiting the Dynamics of Exploitation and Exploration." In *Research in the Sociology of Organizations*, edited by Mikael Holmqvist and André Spicer, 37:161–192. Emerald Group Publishing Limited. http://dx.doi.org/10.1108/S0733-558X(2013)0000037010.

Bamford, S., E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, et al. 2004. "The COSMIC (Catalogue of Somatic Mutations in Cancer) Database and Website." *British Journal of Cancer* 91 (2): 355. doi:10.1038/sj.bjc.6601894.

Bourret, Pascale, and Alberto Cambrosio. 2019. "Genomic Expertise in Action: Molecular Tumour Boards and Decision-making in Precision Oncology." *Sociology of Health & Illness*, June. doi:10.1111/1467-9566.12970.

Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30 (5): 643–683. doi:10.1177/030631200030005001.

Cambrosio, Alberto, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand Jordan, and Pascale Bourret. 2020. "'Overcoming the Bottleneck': Knowledge Architectures for Genomic Data Interpretation in Oncology." In *Data Journeys in the Sciences*, edited by Sabina Leonelli and Niccolò Tempini. Berlin: Springer International Publishing. doi:10.1007/978-3-030-37177-7.

Ebeling, Mary F. E. 2016. *Healthcare and Big Data: Digital Specters and Phantom Objects*. New York: Palgrave Macmillan. http://encore.exeter.ac.uk/iii/encore/record/C__Rb4019293.

Forbes, S, J Clements, E Dawson, S Bamford, T Webb, A Dogan, A Flanagan, et al. 2006. "COSMIC 2005." *British Journal of Cancer* 94 (2): 318–322. doi:10.1038/sj.bjc.6602928.

Forbes, Simon A., David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, et al. 2015. "COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer." *Nucleic Acids Research* 43 (D1): D805–D811. doi:10.1093/nar/gku1075.

Forbes, Simon A., Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, et al. 2011. "COSMIC: Mining Complete Cancer Genomes in the Catalogue of Somatic Mutations in Cancer." *Nucleic Acids Research* 39 (suppl 1): D945–D950. doi:10.1093/nar/gkq929.

Forbes, Simon A., David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, et al. 2015. "COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer." *Nucleic Acids Research* 43 (D1): D805–D811. doi:10.1093/nar/gku1075.

Futreal, P. Andrew, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. 2004. "A Census of Human Cancer Genes." *Nature Reviews Cancer* 4 (3): 177. doi:10.1038/nrc1299.

Green, Sara, Annamaria Carusi, and Klaus Hoeyer. 2019. "Plastic Diagnostics: The Remaking of Disease and Evidence in Personalized Medicine." *Social Science & Medicine*, May. doi:10.1016/j.socscimed.2019.05.023.

Hanahan, Douglas, and Robert A Weinberg. 2000. "The Hallmarks of Cancer." *Cell* 100 (1): 57–70. doi:10.1016/S0092-8674(00)81683-9.

Jubb, Harry C., Harpreet K. Saini, Marcel L. Verdonk, and Simon A. Forbes. 2018. "COSMIC-3D Provides Structural Perspectives on Cancer Genetics for Drug Discovery." *Nature Genetics* 50 (9): 1200. doi:10.1038/s41588-018-0214-9.

Keating, Peter, and Alberto Cambrosio. 2011. *Cancer on Trial: Oncology as a New Style of Practice*. Chicago, IL: University of Chicago Press.

Lee, Francis, Jess Bier, Jeffrey Christensen, Lukas Engelmann, Claes-Fredrik Helgesson, and Robin Williams. 2019. "Algorithms as Folding: Reframing the Analytical Focus." *Big Data & Society* 6 (2). SAGE Publications Ltd: 2053951719863819. doi:10.1177/2053951719863819.

Leonelli, Sabina. 2012. "Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies." *International Studies in the Philosophy of Science* 26 (1): 47–65. doi:10.1080/02698595.2012.653119.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.

Leonelli, Sabina, and Niccolò Tempini, eds. 2020. *Data Journeys in the Sciences*. Springer International Publishing. doi:10.1007/978-3-030-37177-7.

Manovich, Lev. 2001. *The Language of New Media*. Cambridge, MA: MIT Press.

March, James G. 1991. "Exploration and Exploitation in Organizational Learning." *Organization Science* 2 (1): 71–87. http://www.jstor.org/stable/2634940.

Nelson, Nicole C., Peter Keating, and Alberto Cambrosio. 2013. "On Being 'Actionable': Clinical Sequencing and the Emerging Contours of a Regime of Genomic Medicine in Oncology." *New Genetics and Society* 32 (4): 405–428. doi:10.1080/14636778.2013.852010.

Prainsack, Barbara. 2017. *Personalized Medicine: Empowered Patients in the 21st Century?* New York: New York University Press.

Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. "The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers." *Nature Reviews Cancer* 18 (11): 696. doi:10.1038/s41568-018-0060-1.

Tempini, Niccolò. 2015. "Governing PatientsLikeMe: Information Production and Research through an Open, Distributed and Data-Based Social Media Network." *The Information Society* 31 (2): 193–211. doi:10.1080/01972243.2015.998108.

Tempini, Niccolò. 2017. "Till Data Do Us Part: Understanding Data-Based Value Creation in Data-Intensive Infrastructures." *Information and Organization* 27 (4): 191–210. doi:10.1016/j.infoandorg.2017.08.001.

Tempini, Niccolò, and Sabina Leonelli. 2018. "Concealment and Discovery: The Role of Information Security in Biomedical Data Re-Use." *Social Studies of Science* 48 (5): 663–690. doi:10.1177/0306312718804875.

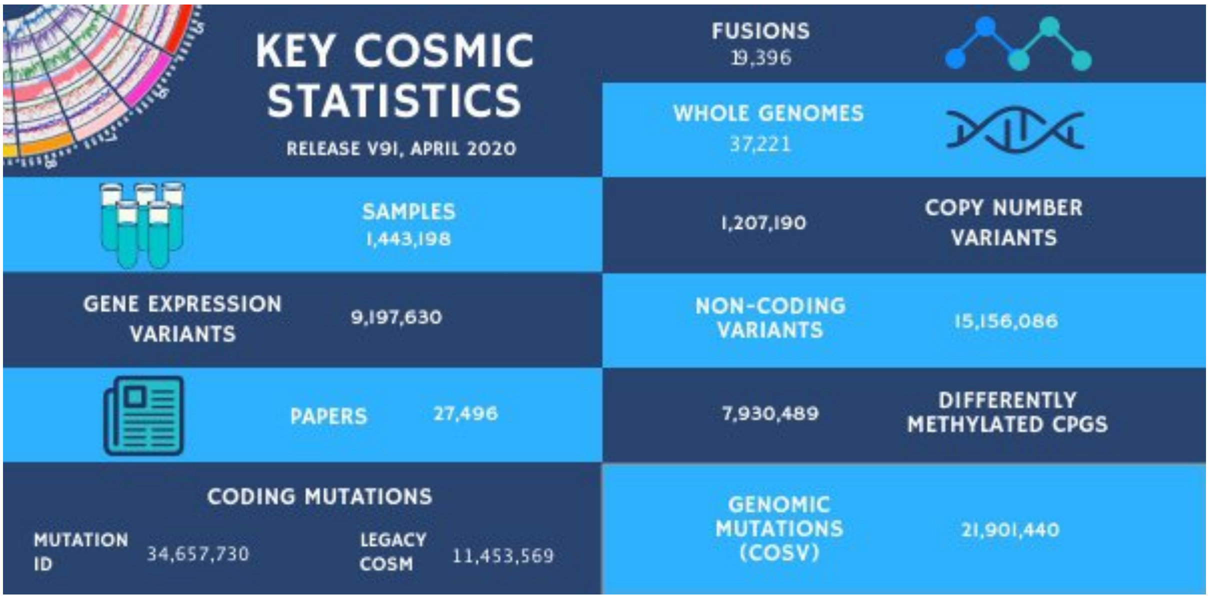
Tempini, Niccolò. 2020. "The Reuse of Digital Computer Data: Transformation, Recombination and Generation of Data Mixes in Big Data Science." In *Data Journeys in the Sciences*, edited by Sabina Leonelli and Niccolò Tempini, 239–263. Cham: Springer International Publishing. doi:10.1007/978-3-030-37177-7_13.

Tempini, Niccolò and Sabina Leonelli. Under review. "Actionable Data for Precision Oncology: Framing Trustworthy Evidence for Exploratory Research and Clinical Diagnostics." *Manuscript under review for international journal.*

Timmermans, Stefan, Caroline Tietbohl, and Eleni Skaperdas. 2017. "Narrating Uncertainty: Variants of Uncertain Significance (VUS) in Clinical Exome Sequencing." *BioSocieties* 12 (3): 439–458. doi:10.1057/s41292-016-0020-5.

Vignola-Gagné, Etienne, Peter Keating, and Alberto Cambrosio. 2017. "Informing Materials: Drugs as Tools for Exploring Cancer Mechanisms and Pathways." *History and Philosophy of the Life Sciences* 39 (2): 10. doi:10.1007/s40656-017-0135-4.

Figures



KEY COSMIC STATISTICS

RELEASE V91, APRIL 2020

SAMPLES 1,443,198

GENE EXPRESSION VARIANTS 9,197,630

PAPERS 27,496

CODING MUTATIONS

MUTATION ID 34,657,730

LEGACY COSM 11,453,569

FUSIONS 19,396

WHOLE GENOMES 37,221

1,207,190 COPY NUMBER VARIANTS

NON-CODING VARIANTS 15,156,086

7,930,489 DIFFERENTLY METHYLATED CPGS

GENOMIC MUTATIONS (COSV) 21,901,440

Fig. 1 Key COSMIC stats from the release of version 91, April 2020 (https://cosmic-blog.sanger.ac.uk/cosmic-release-v91/)

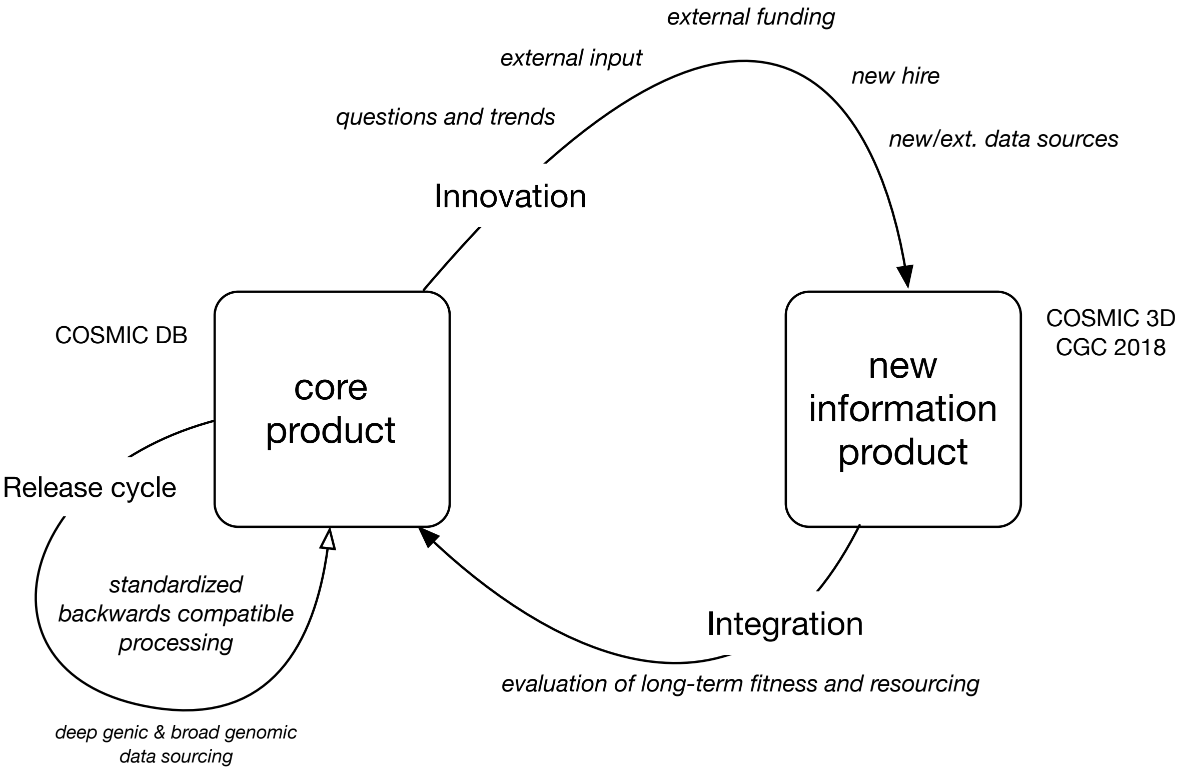


Fig. 2 Summary of COSMIC's evolution patterns

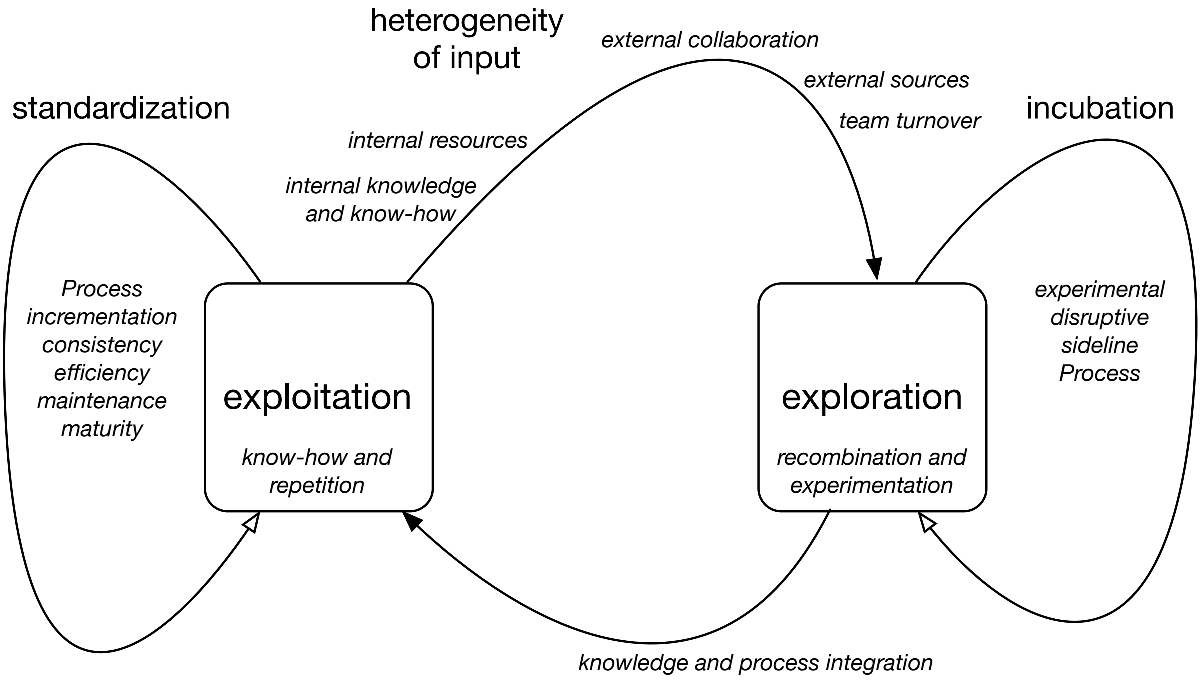


Fig. 3 Organizational learning for sustainable database development

	workforce	**process characteristics**	**adaptive learning**	**database growth**	**release cycle**	**funding stream**	**path to sustainability**
core product	low turn-over	standardization; continuity; efficiency	*exploitation*	extensive, incremental	incremental	stable	continuous development
new product	high-turnover	experimentation; re-combination; project-based	*exploration*	intensive	incubation	project-based	*integration in core product development*

Table 1 Curation activities in COSMIC