Data Observatory

Llamado a Propuestas de Valor

Operations Concept (CONOPS) & Implementation Plan

This document provides information on how does the Data Observatory will accomplish its mission when implemented, and an implementation plan addressing how these operations are ramped up during the first year.

Ministerio de Economía, Fomento y Turismo

Gobierno de Chile

DATAOBSERVATORY

Contents

DO Vision, Mission, and Principles

The DO **vision** is to be at the vanguard of data-centric innovation, leading in the production of data-centric solutions, talent, and social capital for the Latin-American region.

This translates to the **mission** of hosting datasets of global value acquired and generated in the Latin American region, and enabling their maximal exploitation by the global scientific community, the industry, and the public, facilitating data access, analysis, exploration, visualization, and governance.

In accomplishing this mission the DO activities will be compliant with the following principles:

- shall prioritize involvement in fields at the vanguard of data-centric requirements,
- shall foster multi-directional transfer between selected fields and DO-members,
- shall work in coordination with DO-members avoiding competing with them,
- shall aim at complementing DO-members capacities,
- shall promote the use of open and public standards,
- shall promote open access to knowledge,
- shall promote access focused in intended audiences when deploying data,
- and shall be financially sustainable over time.

To fulfill its mission, the DO will organize its activities in 4 lines of work, each further developed in this document: Dataset and Dataset Products, Talent Formation, Challenge management, and Outreach.

In order to provide sustainability to these lines of work, the DO will have a revenue model that has five mechanisms each further developed in this documentl: Founding Members, Memberships, Fees for Challenge Related Activities, Fees for Specialty Datasets, and Grants and Donations.

Lines of Work

The astroinformatics initiative (FIE Grant FIE-2016-V022, CORFO Grant 16IFI6626), was a program lead and financed by the Ministry of Economy, had the mission of identifying and initiating investments and measures in order to foster the digital economy in Chile using the country’s home advantage in Astronomy.

The program generated the basic concepts and structure of the DO. It also delivered a definition of Astroinformatics as a field that is addressing data-centric challenges that lie at the future of the Latin American productive sector, hence creating opportunities to increase Chilean protagonism in astronomy and to foster capacity for the future of digital economy, specifically in data-centric tasks: acquisition & generation, analysis, exploration & visualization, and governance & access of big data (see fig. 1).

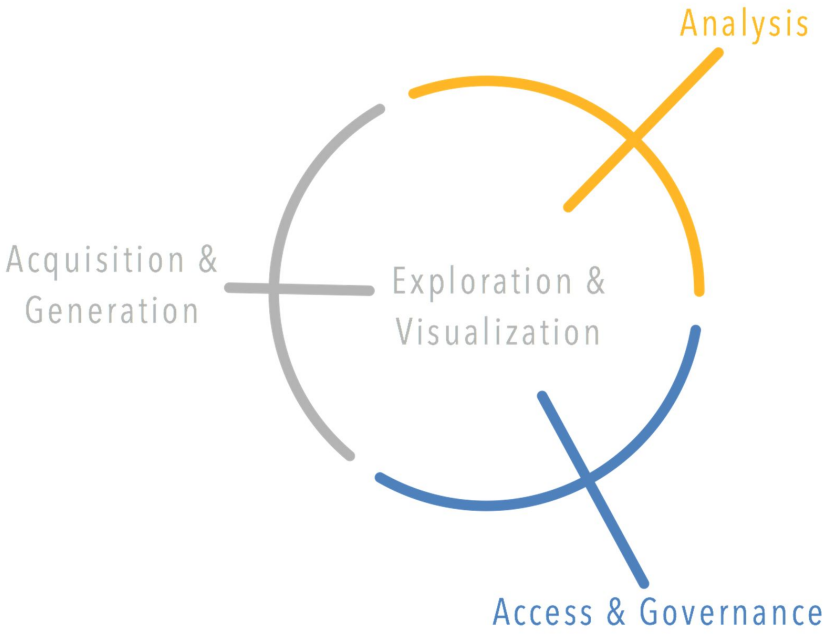

Fig 1. Astronomy Data-Centric Tasks or Astroinformatics. ***Data acquisition and generation tasks:*** *Astronomy projects demand data that can come from astronomical archives, observatories or simulations, the output of them is data that has to be acquired then it has to be converted to appropriate physical units and transferred somewhere for either analysis or curation.* ***Data access & governance:*** *In order to use and reuse the data for research, datasets need to be standardized, stored and indexed, enabling it for search by either the initial scientific project or others that require it, which means that it will be filtered and transferred somewhere else again.* ***Data analysis:*** *Data obtained either from curated archives or directly from simulations or observations is analyzed (either analyzed fast or in a way, not time dominated) in order to obtain new knowledge from it. It is interesting that the process does not end there, and new data and insights may be transferred again to be standardized and eventually made available for search for the use of other projects.* ***Data Exploration and Visualization****: In doing each of these actions, the person explores and visualizes data, and chooses what to do next, it is a non-linear process and it can jump from any point to the next.*

Even though astronomy differs from the productive sector in many aspects, it has similarities in the data-centric tasks. The diagram in figure 1 portrays the tasks that were found to be common between astronomy and mining, precision agriculture, e-health, e-commerce, among other industries in Chile and the region.

The DO lines of work described below will be focused on generating capacities, tools, and methodologies for the development of these tasks.

Dataset and Dataset Products

To fulfill its mission, the DO shall judge dataset ingestion based on the potential of a given dataset to enable DO mission and lines of work. These challenges shall demand state of the art technology and the best available talent.

Astronomical data produced in Chile is the first high-value dataset that the DO will focus on. It will include both observations of the sky from Chile and the data associated with the systems that acquire this data, the most advanced technology for Astronomy. It will always try to inter-operate this data with other open access datasets, hence it will leverage standards creation, and foster use of these standards.

The DO data set will be built based on existing open data sets. However, curated versions for specific purposes such as research, talent formation, algorithms testing, and others will be made available for a fee. Each one of these subsets of the data set will be made available through a series of tools to be created by the DO, or as part of DO challenges as described below.

The DO data set will be cloud-based, using technologies like containerization and serverless computing, and potentially federation of storage and computing resources from contributing members. The DO will allow the use and reuse of the data by commercial cloud companies under yet to be defined access policy and governance.

The data set will include all the original metadata, plus DO-specific derived metadata to help with the different scientific and operational goals of the DO.

Activities for the generation of the data set include:

- generation and maintenance of agreements with observatories in Chile for a constant transfer of data to the DO; this includes maintaining ingest pipelines for the periodical addition of new data;
- identification and capture other relevant open data sets that can be of value to the DO mission;
- processing, storage, and curation of the DO data set following interoperable standards and complying with open data frameworks.

The diagram below summarizes the different possible subsets of data and access model

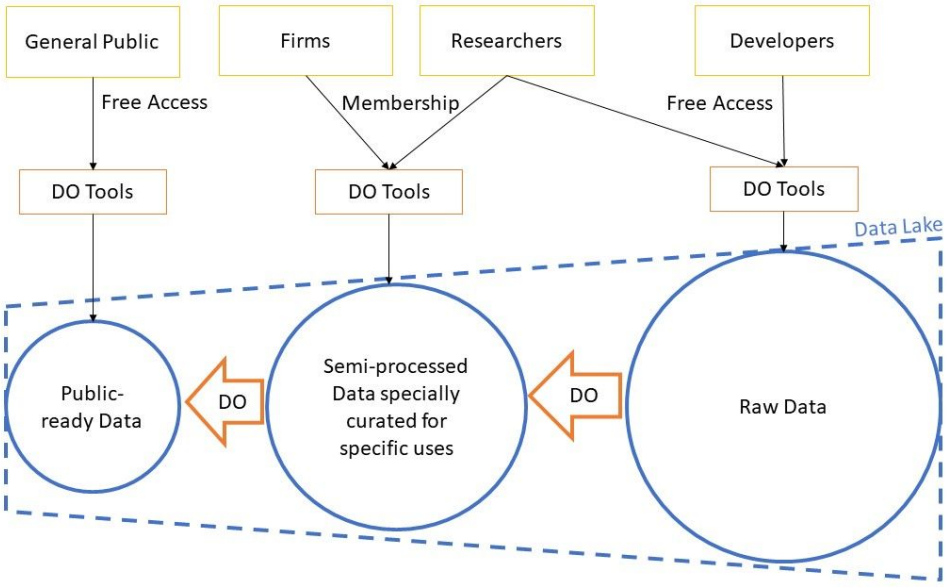

Fig 2. Data Observatory data-access models.

Public-ready data set definition

The DO will generate a highly processed data set, to be accessible through DO and partner generated tools, free of charge to the entire community. This data set and tools will focus on outreach activities and will aim to promote data science, astronomy, and curiosity to all interested in the sky and the universe. It will also be available through Virtual Observatory protocols and standards.

Semi-processed data set definition

Different activities required different structures, sizes, and characteristics. The DO will generate a series of data set destined to:

a) train professionals in the activities related to data
b) test and train algorithms for different industrial functions
c) support research in astronomy and other disciplines studying physical phenomena

Access to this data set will take place in the form of memberships as described below, always promoting open data frameworks for non-commercial activities.

Raw data definition

The DO will create and maintain a raw data set with the data as ingested from different sources according to the agreements the DO enters into. Such data set will replicate legal and contractual access conditions from the source. However, the DO will facilitate technical access conditions, to the largest possible public by the use of

the cloud. In particular, datasets from astronomical observatories will also be accessible through VO protocols and conform to VO standards.

Associated Computing
In order to provide tools to access, analyze, visualize, and explore the DO data for non-commercial uses, the DO members will have access to cloud credits depending on the agreements reached between the DO and Infrastructure-as-a-Service founding and regular members. These services will be paid to the DO in a way compliant with the stakeholder's procurement guidelines, and the amount paid and the provision of the service will occur according to the agreements reached between the DO and Infrastructure-as-a-Service founding and regular members.

Talent Formation

The DO will collaborate with talent formation offer stakeholders (academy, open, others) and productive sector stakeholders (industry, culture, government, others) with hands-on direct learning opportunities related with challenges in the mentioned areas, related with the use of the DO data set, the DO partners needs, and constant interaction with the DO stakeholders.

An initial exploration study was performed in order to identify specific opportunities for the DO and it is available here. According to what was found, the DO should focus on the Access and Governance that was found not to be covered thoroughly in the different educational programs.

In accomplishing this mission, the DO will work towards:

- Complementing capacities of formal educational institution: The DO will support, in a first stage, initiatives like "La Serena School for Data Science (LSSDS). The DO will build on that world-leading learning experience and expand its impact on wider audiences, based on the needs of productive sector stakeholders. Specifically, the DO could add value with Access and Governance courses.
- Consolidating talent formation efforts such as LSSDS, complementing with professional master programs and academy fellowships. These programs will not be focused on creating talent for astronomy, but for the productive sector stakeholders, with the need for data professionals prepared for their acquisition, analysis, visualization, and archive challenges. The DO will not grant professional degrees but will work in close collaboration with Universities in this area providing courses in Access and Governance, challenges for master or Ph.D. theses and access to the datasets.
- Generating Certifications Programs: As a joint-venture with partners, the DO will provide certification programs related to the DO or a partner's product or

activities. The DO will generate contents using partners platforms for its distribution.
- Generating tools like games, tests and other teaching materials for teachers and professor in all educational level.
- Generating content material to be shared through MOOC, EDx or other existing platforms. This courses should focus on Access and Governance or, if done in more general data science skills, be in Spanish.

Challenges management

In performing its mission as a neutral-broker, the DO will organize its research, development and innovation activities in the form of Challenges. Challenges will arise from the needs of stakeholders or directly from the DO.

DO's generated Challenges: In order to accomplish its mission, the DO will be constantly seeking for challenges that enable its participation in the forefront of data-centric innovation. These challenges shall demand the most advanced technology and the best talent available. The DO will foster the creation of challenges that comply with said characteristics and can be solved through the DO data-set, products or people, or its network's. Challenges will also have to serve the needs of one or more of the DO stakeholders. Challenges will also aim at involving the startup community in DO related activities.

Stakeholder generated Challenges: The DO will evaluate challenges formulated by researchers, observatories, firms, the public sector or any organization and will create a pipeline of challenges through which they will be tackled. Evaluation criteria TBD but linked with the mission.

DO participation in each challenge will array from the mere coordination to the development of solutions, all in collaboration with DO stakeholders as described below:

1. **Brokerage:** The DO will identify and match individuals or organizations with the capacities to solve an identified challenge. This could be done through a direct match or an open innovation contest.
2. **Services:** The DO or a Stakeholder will carry-out support and consulting activities in:
    2.1. the definition of challenges to be resolved using DO or Stakeholder assets including data and/or products,
    2.2. the project management and systems engineering required to face the challenge by a team involving diverse stakeholders,

2.3. expertise in product improvement, deployment, operations, and disposal.

3. **Product Development:** The DO or a Stakeholder will develop solutions that will then be licensed by itself or by a Stakeholder (hence DO will be licensee). The licensing format should promote open-source and open-data models. Licensing terms will be defined for each challenge depending on the DO and stakeholders involvement.

Outreach

Astronomy is at the heart of Chile's culture and identity having a good level of knowledge and remembrance in the population, as an “Imagen de Chile” survey (2016) shows with results like that the 87% think Chile is recognized as a privileged place to observe the universe, an 84% think Chile can be recognized for having the best sky of the world, and a 77% think Chile is a renowned place for the installment of observatories. Finally, the Chilean sky is considered the fifth more representative element of the country, as important as poetry that has two Nobel prizes, and above soccer the year after winning the America Cup.

The DO will further promote education and culture based in astronomy aiming towards data-centric education and talent formation, developing outreach activities for national and international stargazers, data enthusiasts, children, and the general public. As examples, such activities may include:

- Generating human-computer interfaces for space immersion. Tools to “feel like” in outer space, including video games and others.
- Web 2.0 content about data science and astronomy as economic development tools.
- Supporting K12 institutions with education-material, school contests using astronomical data, data science in astronomy, and science in general based on the knowledge of physical phenomena developed by astronomy
- Hosting conferences and seminars about data science and astronomy, virtual observatory, astroinformatics, astronomy for development.
- Coordinating an international astroinformatics summit to be held annually in parallel to the Data Observatory board and advisory committees meetings.
- Generating Data Science Contests related to astronomy to promote citizen science.

Value Proposition for DO Stakeholders

These line of works will create a spectrum of value for a diverse set of stakeholders:

For **Infrastructure as a Service Providers (Cloud and others)**, it will provide an attractive pool of data to be hosted at their infrastructure, this data will increase their

infrastructure associated tools usage and brand value, and through the DO challenges with diverse stakeholders, it will increase the spectrum of potential customers on one hand and talent for the provider on the other. Finally, it will broaden the provider impact, specifically to the area of talent, technology, and infrastructure capacity development.

For **Data Producers**, it will give untethered access to their public data, and their preferred specific-audiences access to other datasets they provide to the DO. It will enable collaboration with a large and active data science community, boosting the productivity of their data in a broad sense that includes research publications and can lead to talent, technology and infrastructure development for the Latin American region. Finally, through the DO challenges line it will provide new tools to exploit their data with DO solutions at TRL6 to TRL9, including potential royalties if these tools were generated using their data.

For **DO Trainees**, it will provide untethered access to all public data in the DO, including access to specially curated datasets for education and support materials and specific research needs, and it will provide preferred Data Producer access to other datasets. It will also provide opportunities to develop relevant Data Science skills and to participate as a protagonist in challenges that are defining the future of the data science field and the data-centric activities in a broad spectrum of domains. Finally, it will provide increased domain exposure, to the domains of the DO data producers.

For **Small, Medium-sized Enterprises (SMEs) and Large Enterprises**, if they have data-centric capacities, it will provide opportunities to scale up these capacities in volume, variability (complexity) or velocity to serve new and more profitable markets, and no matter if they already have capacities or not, it will provide data-driven capacities (talent formation environment, solution development environment), skilled and experienced Data Professionals to hire for projects or indefinitely, will provide ways to evaluate the commercial viability for developed solutions and royalties if the solutions are successful. Finally, it will increase their brand value.

For **Research Organisations**, it will provide untethered access to public data, and access to curated datasets for education and support materials or solving specific research challenges, it will also provide skilled and experienced Data Professionals, interaction with domain experts, and commercial opportunities for researched

solutions, including potential royalties. For these organization, it will also provide opportunities to increase brand value and to broaden the organization impact.

For **Researchers**, it will provide untethered access to public data, and access to specially curated datasets for research challenges, a pool of skilled and experienced data professionals, and interaction with domain experts, it will also provide commercial opportunities for researched solutions and potential royalties.

For **Higher Education Entities**, it will provide untethered access to public data, and access to specially curated datasets for education and support materials, training materials and tutors.

For **Schools**, it will provide untethered access to public data and access to specially curated datasets for education and support materials, training materials and tutors, a place to obtain hands-on peta-scale data science and engineering training.

For **Data-gazers**, it will provide untethered access to public data, a place to obtain hands-on peta-scale data science and engineering training and philanthropy opportunities.

Operational Principles

In order to produce successful solutions or systems through each line of work, the DO will follow the Systems Engineering perspective with a strong orientation to agile best practices (this means to develop the systems and solution in an iterative and recursive fashion, much more than in the classical space-systems waterfall fashion). The figure below is commonly known as the V-diagram and provides a high-level overview of the Systems Engineering process. It can be explained either in abstract terms or with the examples in the footnote (see footnote[1])

[1] **A new dataset for the "Dataset and Dataset Products line"**: a given domain-stakeholder has needs that can be solved by acquiring a dataset, such as a planet formation researcher interested in learning about the statistics of protoplanetary disks in our universe. From the system perspective, this translates to several "requirements", to illustrate we name 3: (1) to acquire data to form a dataset to be analyzed, (2) composed of observations of protoplanetary disks in diverse wavelengths (3) suitable for analysis in terms of data interoperability. These requirements can be translated to a "specification" by selecting exactly what features of the protoplanetary disks will be analyzed, defining which telescopes to use, establishing how will the telescopes will calibrate and deliver the data, and how the data will be structured to answer the stakeholders need.

A new product from the "Challenges Management line": after successfully finishing a challenge management activity, a new product or solution is created, the "performance requirements" that the product achieves can be seen as potentially satisfiable stakeholders needs, and is possible to identify fields with stakeholders having these needs, in order for the DO to associate with commercial partners for a technology transference challenge with these identified fields as targets.

A new learning material for the "Talent formation line": a certain university degree program has the need of providing hands-on skills and experience in finding optimal ways of performing noise removal for large-volume datasets. This translates to several requirements: (1) to have a dataset useful for the task; (2) to enable access to work with this dataset for the researches to find the methods they seek. These requirements can be translated to a specific solution, or "specified" by selecting exactly where to

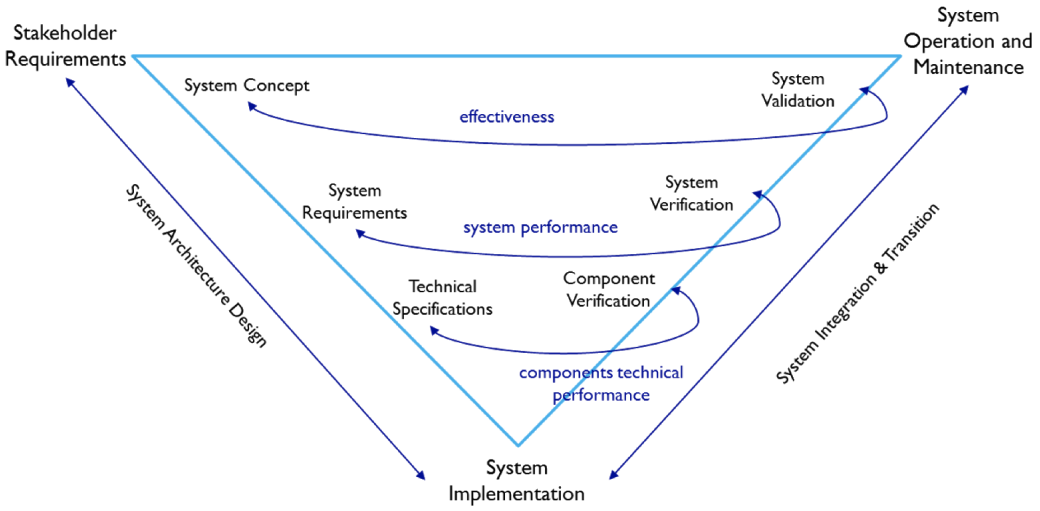

Figure 3 the: Systems Engineering process or V-diagram.

In order to explain the diagram from a **top-down perspective**, the processes go from the domain perspective or "stakeholder perspective", to the perspective of the DO solution or system. The DOers will synthesize a "System Concept" eliciting and analyzing "Stakeholder needs" and will transform it to performance characteristics or "System requirements" that the DO products and solutions shall do to satisfy these needs. Selecting the best specific solution that achieves the required performance generates "system specifications". This last step enables to architect solutions and later to implement them, integrate them, and verify that the integrated sub-systems and eventually the System or solution is compliant with the performance required and when the System (tool, solution or product) is transitioned to operations, the Stakeholders can validate it satisfies the needs they have.

In order to explain it from a **bottom-up perspective,** the processes help derive a System Concept from the technical specifications of a solution that is known possible. This is useful in case alternative solutions are going to be sought, System Requirements could be derived too for processes to help in the search of candidates that might be better than current specific solutions to stakeholder needs.

store the data, in what notebook to program the tutorial, and how to deliver the material on a given degree course.

Data Observatory Staff

In order to perform the activities derived from each DO line of work, the DO will have the following core competencies, responsibilities per line of work and organizational structure.

Core Competencies

Table 5: DO Core Competencies

Core Competency	Requirements
Leadership DOers	The DO shall have people with strong expertise in leadership, project management and systems engineering, and a demonstrated record of leadership of multi-disciplinary and multi-cultural groups delivering innovation.
Domain DOers	The DO shall have strong domain experience and expertise, and the ability to know how does the domain operates and provide insight into stakeholder needs that can be solved with data-centric systems.
Data DOers	The DO shall have strong engineering expertise in areas related to data acquisition & generation, analysis, exploration & visualization, access & governance of data.
Engineering DevOps DOers	The DO shall have experts to develop and operate for each of the computing architecture subsystems, including cloud and HPC computing, systems architecture, databases, specific frameworks and software TBD.
Talent DOers	The DO shall have education experts including at least a person per target audience. It shall have strong and vast data-centric talent formation expertise, aimed at least to the following audiences: from children to high-school and from undergrad, graduate to professional.
Business Development DOers	The DO shall have experts in all stages of product, solution and business development lifecycle (end-to-end development), from concept to deployment to operations.
Outreach and Engagement DOers	The DO shall have strong experience in strategic-communication, organization of outreach activities and divulgation to both general and specific audiences, aimed at least to the following audiences: data producers (astronomical observatories and other DO members initially), public sector, private sector, academy and curiosity-driven people interested in the DO data such as stargazers, datagazers, and explorers.
Administration, Finance and Legal DOers	The DO shall have strong legal, administrative, and financial management expertise.

Organizational Structure

The DO organizational structure, portrayed in figure 4, will operate as a matrix structure, combining domain and specific capacities, with the intersection being every DO Line of Work. Domain capacities will be tied to the specific activities that originate the DO datasets, and will drive the understanding of how to generate data-driven value for the domain. Horizontal capacities will rely on the core competencies and will have a strong mandate of a permanent focus in the Domain needs, in all Lines of Work. Such interactions will be performed following the System Engineering perspective described in figure 3, with a strong agile orientation, as said before.

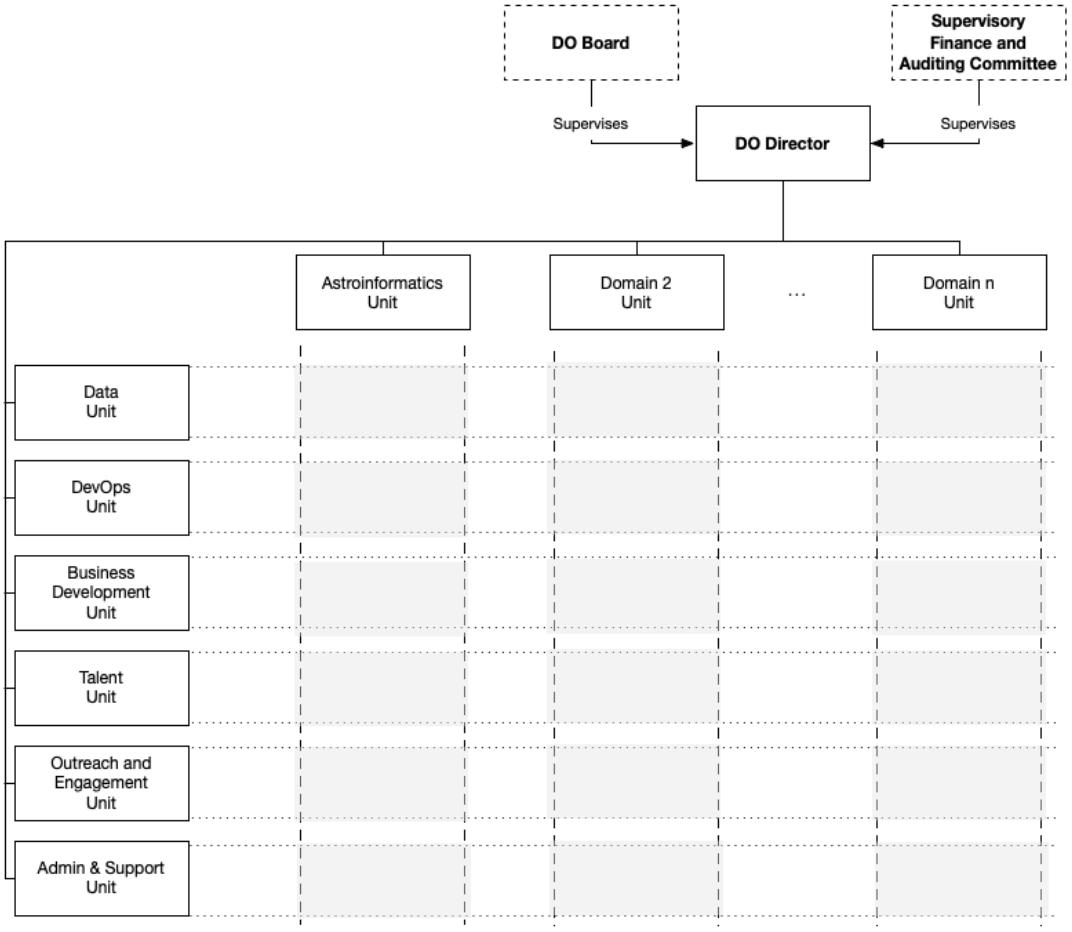

Figure 4. Data Observatory Organizational structure

Computing Architecture and Interface with Data Sources

The intended architecture for the Data Observatory is as a cloud-native infrastructure which allows for multi-vendor/multi-provider usage. The main advantages of the proposed architecture are the inherent horizontal and vertical scalability of cloud-native solutions, the ability to avoid single-vendor lock-in, and the capability of integrating local resources, or resources from other international partners, into the system. By means of abstraction layers to the access to processing and storage resources it is possible to start small, with a single provider, and later increase the complexity of the system for the multi-vendor, multi-provider resources.

Figure 5 shows a layered view of the proposed DO architecture. A detailed description of the computing architecture is described in Annex 2.

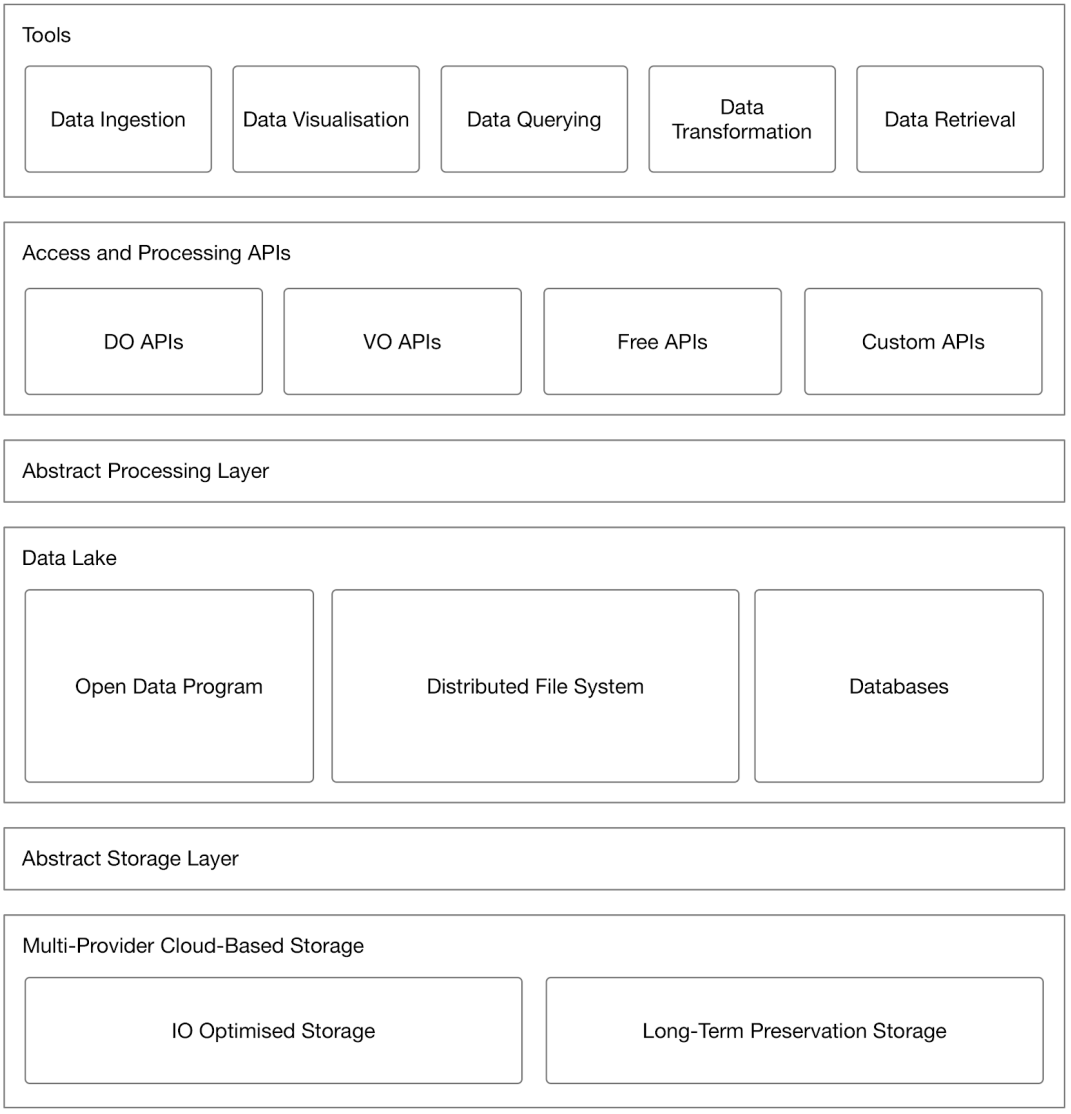

Figure 5: Layered View of the DO computing resource architecture.

Revenue model

To finance its operation, the DO has identified revenue streams described below.

Founding members - Governance

The DO will have 4 founding members, one of them being the Chilean Government. Founding members are those that contribute to the DO in an amount equal to or greater than the Chilean government initial investment, whose membership length will be agreed on a case by case basis, using the Chilean government case as a baseline. DO founding members will be part of the DO governance with the Chilean government according to the diagram below and will have, in addition, the following rights:

- designate ⅗ of the board and ⅔ of the supervisory and auditing committee,
- access to the DO curated data sets and DO-generated challenges,
- access to all DO outreach activities,
- and when Challenge management activities result in a product that will be commercialized by the DO Stakeholder, participation in the agreements that will be generated to correctly assign IP results.

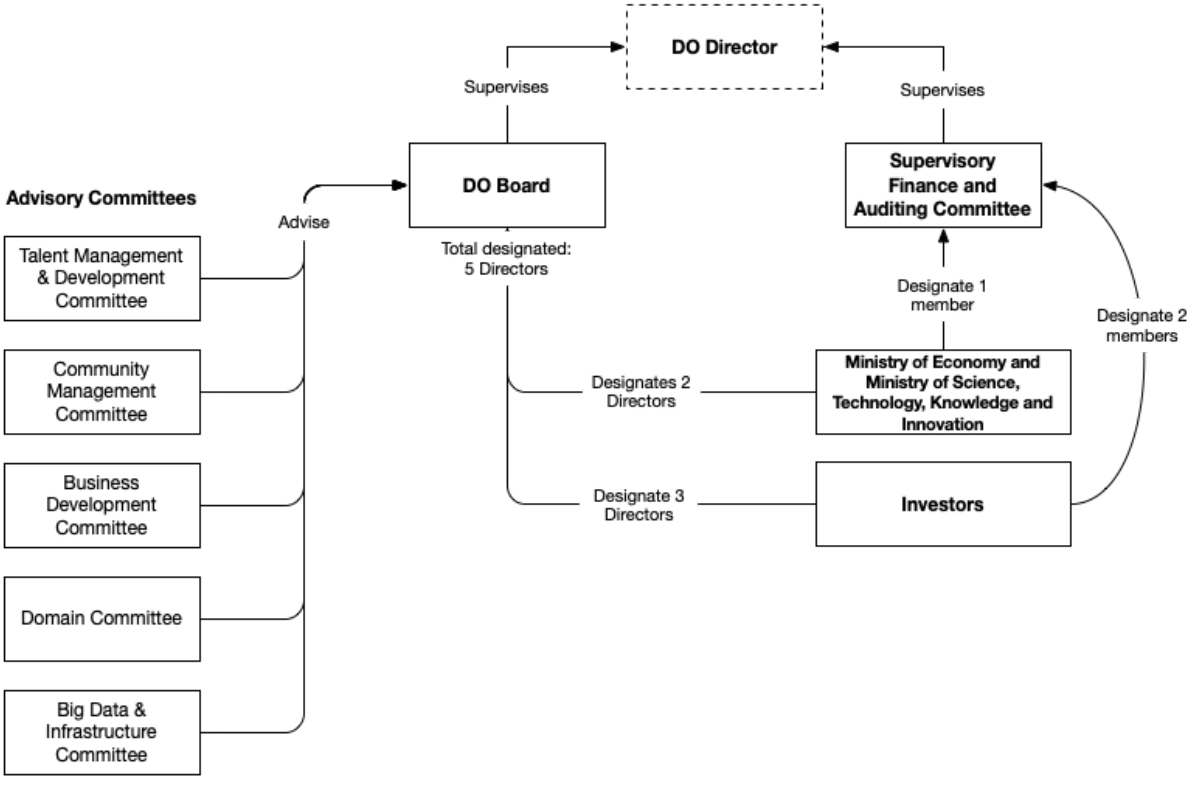

Figure 7. Data Observatory Board Composition. The DO Board will be composed of 5 members designated by the Government (2) and the Founding Members (3), the committees will provide expert advice to the board on the mission of the DO.

Memberships

There are a number of ways to become a DO member:

1. being a DO stakeholder that pays an annual membership-fee, varying based on the size of the entities and the entity mission;
2. being a DO stakeholder that provides a dataset deemed valuable by the DO in accordance with its mission. Data is considered valuable when associated challenges that enable the DO to be at the vanguard of data-centric activities can be identified. The length of the membership, in this case, will be the period over what data is contributed and remains valuable.
3. being a DO stakeholder that provides cloud capacities deemed valuable by the DO in accordance with its mission. Cloud capacity is considered valuable when enables addressing DO challenges.
4. being a DO stakeholder that provides talent deemed valuable by the DO in accordance with its mission.
5. the DO board can define new ways to become DO member.

DO Members will have the following rights:

- designate members of the advisory committees,
- access to the DO curated data sets and DO-generated challenges,
- access to all DO outreach activities,
- and when Challenge management activities result in a product that will be commercialized by the DO Stakeholder, if the member is included in the Challenge, participation in the agreements that will be generated to correctly assign IP results.

Fees for challenge-related activities

Non-DO members will be able to participate in DO generated challenges as well as propose challenges in exchange of a brokerage fee. In case the DO participation includes technical resources from the DO and/or generate a new product that will be commercialized by a Non-DO or a DO member, special agreements will be generated to ensure the correct IP structure and derived royalties for the DO.

Fees for specially curated data sets

Non-DO members will be able to access specially curated data sets for a fee. Such fee will vary depending on:

- the size of the entities
- the entity mission

Grants and donations

In order to fulfill its outreach and public interest activities, the DO will apply for grants for the promotion of culture, education, science, and innovation. Also, the DO will also receive donations and work with the local industry, particularly those operating in the north of Chile, to channel resources from social responsibility to outreach related activities

Key Performance Indicators (KPIs)

The DO's KPIs are being developed to ensure that it fulfills its mission through its four lines of work. To do so, the KPIs were divided into five groups: general, data lake and data lake products, challenges management, talent formation, and outreach. In each of the lines of work, the KPIs were divided into different groups when needed (for example, KPIs for data lake and data lake products include KPIs related to the data lake, to the scientific production and to the products developed). Also, each of the KPIs is related to one or more of the eight mandates of the DO (see page 2, the "shall…" phrases) to ensure their fulfillment. Finally, relevant stakeholders (from the list available on page 31) were related to each of the KPIs.

A final version of the KPIs is in the process of being developed. Timelines associated with the KPIs will be defined based on the DO budget, after the founding process is complete. For reference purposes, the table below includes some of the KPIs created. KPIs for all lines of work are included in Annex 3 of this document.

General KPIs

These KPIs are not related to a specific line of work of the DO, but to its general operation. Their objective is to ensure that the DO is sustainable.

KPI	Definition	Stakeholders	Mission
Gross Margin	Income - Costs	DOers	8
Public Funding	The share of the incomes that come from public funding	DOers and Government	8
Investors' Funding	The share of the incomes that come from private investors	DOers and Large Enterprises	8
Operational Income	The share of the incomes that come from Memberships and other activities performed by the DO.	DOers, Large Enterprises, SMEs and Government.	8
Donors Funding	The share of the incomes that come from donors.	DOers	8
Cash-In kind Ratio	The share of the incomes that are in-kind contributions.	DOers	8
Growth	Incomes (n+1) / Incomes (n)	DOers	8

Implementation Plan

As part of the DO concept of operations, the DO team is developing an implementation plan that aims at organizing and prioritizing the first year of operation considering a minimum investment that includes the Government contribution and three additional founding members. Pending a final version of the implementation plan, that includes milestones in the areas of operations, computing, staff, communications and risk management, this document includes the key milestones for starting operations and the dates they need to be fulfilled in order to ensure start up of the organization.

1. Legal Milestones

The DO will be created as a Non For Profit Organization according to Chilean Law, specifically as a “Fundación”. A Fundación is a NPO created by group of assets that are put together and exploited towards an agreed goal. In the case of the DO, the assets are those contributed by the founding members, and the Government ability to coordinate collaboration with observatories based in Chile. The legal requirements for the DO constitution include the signing of the bylaws, and the first board session. According to the DO governance model, 5 advising committees need to be established to advise the work of the board.

Activity	**Deadline**
Signing of the bylaws	April 2019
First board meeting	April 2019
First advisory committee sessions	January 2020

2. Administrative Milestones

The DO operations are carried out from the Ministry of Economy by a team that combines Ministry´s employees and FIE advisors. This team will transition to the DO once it is created, with the specific task of founding the organization and ensuring the minimum requirements for a successful operation are in place. Part of these tasks include a national and international search for a DO executive director. The team will continue to operate from the Ministry of Economy until a permanent office is defined.

Activity	**Deadline**
Call for DO executive director	September 2019
Full contractual transition of the current DO team to the DO	December 2019
Offices inauguration	November 2019

3. Revenue milestones

The DO revenue model relies on investors participation, membership fees, royalties for the commercialization by DO stakeholders of products created with DO participation, and fees collected from services provided by the DO. During the first year of operation the DO will secure participation from stakeholders in three categories: DO founding member, DO member and non-members.

Activity	**Deadline**
Agreements for 3 DO founding members	April 2019
Agreements for 3 DO regular members	November 2019
License - IP agreement with 3 DO non members	December 2019

Annex 1: Staff Responsibilities per line of work

The following annex describe the core responsibilities of the DO staff in order to fulfill the DO mission and vision.

For the systems and solutions developed as results of each line of work activities, the responsibilities will be assigned driven by the core competencies of the DOers and the critical tasks of each line of work.

Critical tasks of DOers during Architecture Design of DO Solutions

Table 6: DOers and their ultimate responsibilities per Line of Work during Architecture Design activities

Skill / Line of Work	**Systems pertaining to Dataset and Dataset products**	**Systems pertaining to Talent formation**	**Systems pertaining to Challenges Management**	**Systems pertaining to Outreach**
Domain DOers	elicit and analyze domain stakeholder needs and derive key performance requirements from them[2]			
Data DOers	derive and analyze DO acquisition, analysis, visualization, and access & governance system requirements[3]			
Engineering DevOps DOers	derive and analyze DO DevOps requirements and analyze other requirements from a DevOps perspective. derive and analyze computing technical specifications.[4]			
Talent DOers	elicit and analyze DO audiences needs, derive and analyze DO Talent requirements and analyze other requirements from a Talent perspective.			
Outreach and Engagement DOers	elicit and analyze DO philanthropic and members (current and to be) needs, derive and analyze DO engagement requirements and analyze other requirements from an engagement perspective.			
Business Development DOers	derive DO Business Development requirements and analyze other requirements from Business Development perspective			
Admin, Legal and Finance DOers	analyze all requirements from Admin, Legal, and Finance perspective.			
Leadership DOers	Manage the work, interfaces and balance the team for success			

Critical tasks of DOers during Implementation of DO Solutions

Table 7: DOers and their ultimate responsibilities per Line of Work during Implementation activities

Skill / Line of Work	**Systems pertaining to Dataset and Dataset products**	**Systems pertaining to Talent formation**	**Systems pertaining to Challenges Management**	**Systems pertaining to Outreach**
Domain DOers	Verify implemented systems comply with key performance requirements.			
Data DOers	Implement data systems. Verify implemented systems comply with the acquisition, analysis, visualization, access requirements			
Engineering DevOps DOers	Implement computing systems. Verify implemented systems comply with technical specifications and DevOps requirements.			
Talent DOers	Verify implemented systems comply with Talent requirements.			

[2] I.e., they understand what it takes to be successful in the domain (astronomy initially)

[3] I.e., they understand what it takes to build a successful data system given domain needs

[4] I.e., they understand what it takes to architect and implement a successful system in the cloud

Outreach and Engagement DOers	Verify implemented systems comply with Engagement requirements.
Business Development DOers	Verify implemented systems comply with Business requirements.
Admin, Legal and Finance DOers	Manage admin, legal and financial aspects of implementation.
Leadership DOers	Manage the work, interfaces and balance the team for success

Critical tasks of DOers during Integration and Transition to Operations of DO Solutions

Table 8: DOers and their ultimate responsibilities per Line of Work during Integration and Transition to Operations

Skill / Line of Work	**Systems pertaining to Dataset and Dataset products**	**Systems pertaining to Talent formation**	**Systems pertaining to Challenges Management**	**Systems pertaining to Outreach**
Domain DOers	Validate implemented systems satisfy stakeholder needs.			
Data DOers	Operate data aspects of the system			
Engineering DevOps DOers	Operate engineering DevOps aspects of the system			
Talent DOers	Validate implemented systems satisfy audiences needs. Operate talent aspects of the system			
Outreach and Engagement DOers	Validate implemented systems satisfy philanthropy, founding, and regular DO member needs, Operate Engagement aspects of the system			
Business Development DOers	Operate business aspect of the system			
Admin, Legal and Fin DOers	Manage admin, legal and financial aspects of the transition to operations.			
Leadership DOers	Manage the work, interfaces and balance the team for success			

Annex 2: Computational Architecture

Figure 5 shows a layered view of the proposed DO architecture. A detailed description of the computing architecture is described in Annex 2.

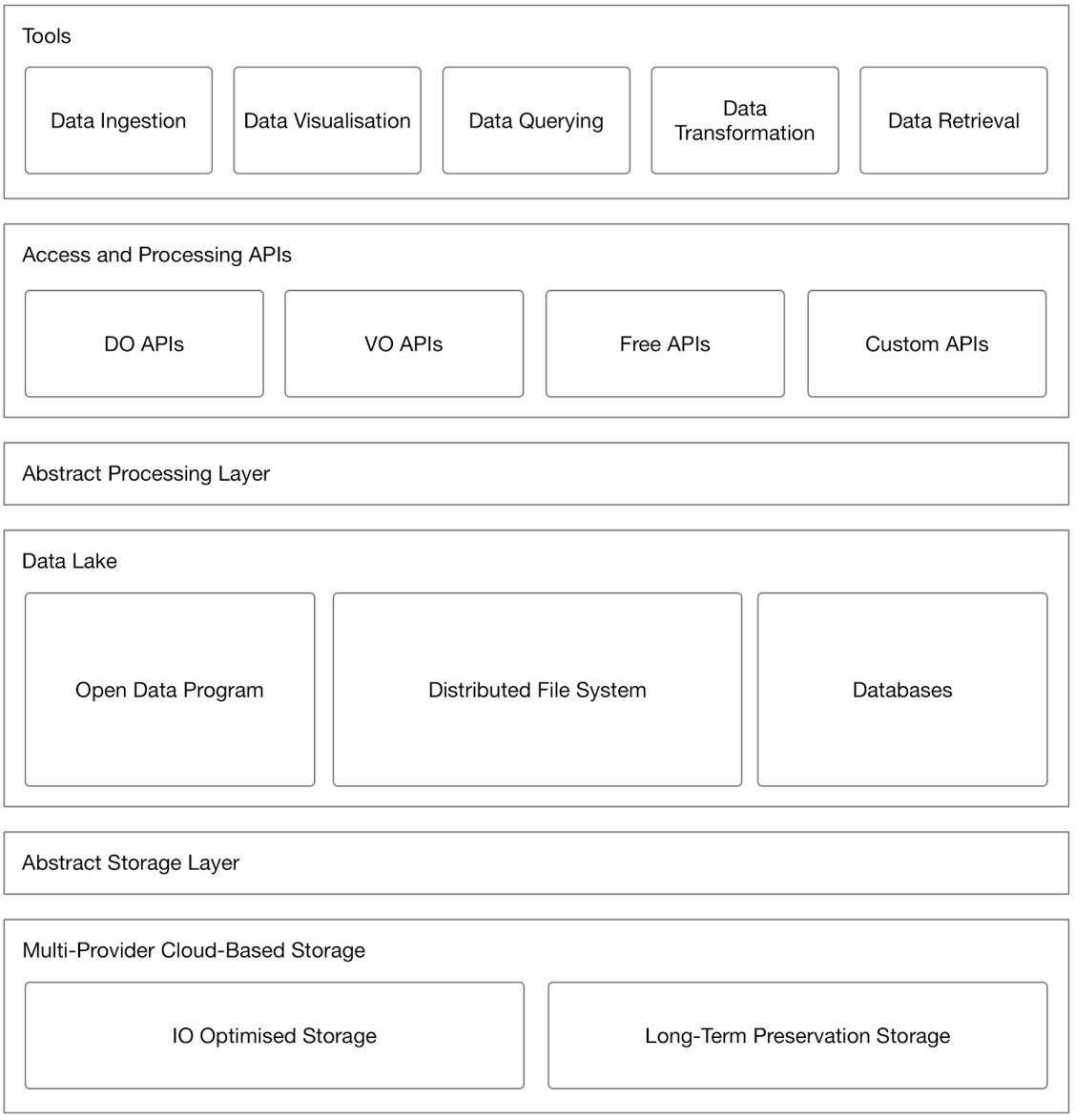

Figure 5: Layered View of the DO computing resource architecture.

In Figure 5 a number of layers are identified:

- **Multi-Provider Cloud-Based Storage:** this bottom layer provides storage services both for Input-Output Optimized Storage —i.e., storage with stringent latency and throughput requirements—, and for Long-Term Preservation storage —optimized for cost per petabyte—.

- **Abstract Storage Layer:** to be developed by the DO, this layer isolates the differences between different kinds of storage, and the different cloud and/or storage resource providers. That layer will support either data replication and redundancy —easier to implement, but potentially less powerful—, or full abstraction of distributed block storage across the different providers —more difficult to implement, but allows for using several cloud and infrastructure storage providers as seamless, redundant, failsafe storage—.
- **Data Lake:** the data lake layer is a logical construct that can support both the operational data for the DO —data ingested from observatories, data generated by users while manipulating DO data, dataset products, and many others—, and datasets which are specifically named as Open Data. The Data Lake will provide a distributed file-system view to the upper layers. We're also including logically in this layer the different databases —either relational or NoSQL— that support products such as catalogs, and multiple indexes to the data. Some additional components in this layer are still TBD in order to support advanced data products such as catalogs, and alerts.
- **Abstract Processing Layer:** to be developed by the DO, this layer isolates the different computing resources that can perform processing tasks on top of the DO data lake, be them from commercial cloud providers, or from DO-dedicated local resources in Chile. DO monitoring information is also generated in this layer and stored in the Data Lake.
- **Access and Processing APIs:** this layer is a collection of Application Programming Interfaces that further abstract the processing into particular functions. Currently we have divided them into four kinds: a) DO APIs —for instance, those that require authentication and authorization—; b) Virtual Observatory (VO) APIs —such as Simple ConeSearch, Simple Image Access v2, Simple Spectral Access v2, Table Access Protocol v1.1, ObsCore, and Datalink, at least—; c) Free APIs —i.e., those non-VO APIs that don't require DO membership; and d) Custom APIs that can be potentially developed for particular tools, purposes, and/or customers.
- **Tools:** the topmost layer, built atop the Access and Processing APIs, collects the tools that we currently foresee for at least Data Ingestion, Data Visualization, Data Querying, Data Transformation, and Data Retrieval functions.

The above architecture has only one point of dependency with respect to the astronomical domain, which is the explicit support for VO APIs.

On top of that architecture, we've devised a functional view for how the Data Observatory will support use cases for astronomy which is shown in Figure 6.

Multiple sources of data —namely, observatories with an agreement with the DO— provide multiple streams of data that have to be ingested into the DO. The ingestion consists not just on the reception, and potentially decompression of the data streams from the observatories, but also relevant metadata extraction, trying to support at least ObsCore, but potentially deeper levels of the IVOA Characterisation Data Model. The data and metadata are logically contained in the DO Data Store, which corresponds to the Data Lake layer in Figure 5, while the ingestion processes correspond to the DO Ingestion tools in the same figure.

The DO will support ingesting advanced data products as well; this includes, but is not limited to, catalogs and alerts.

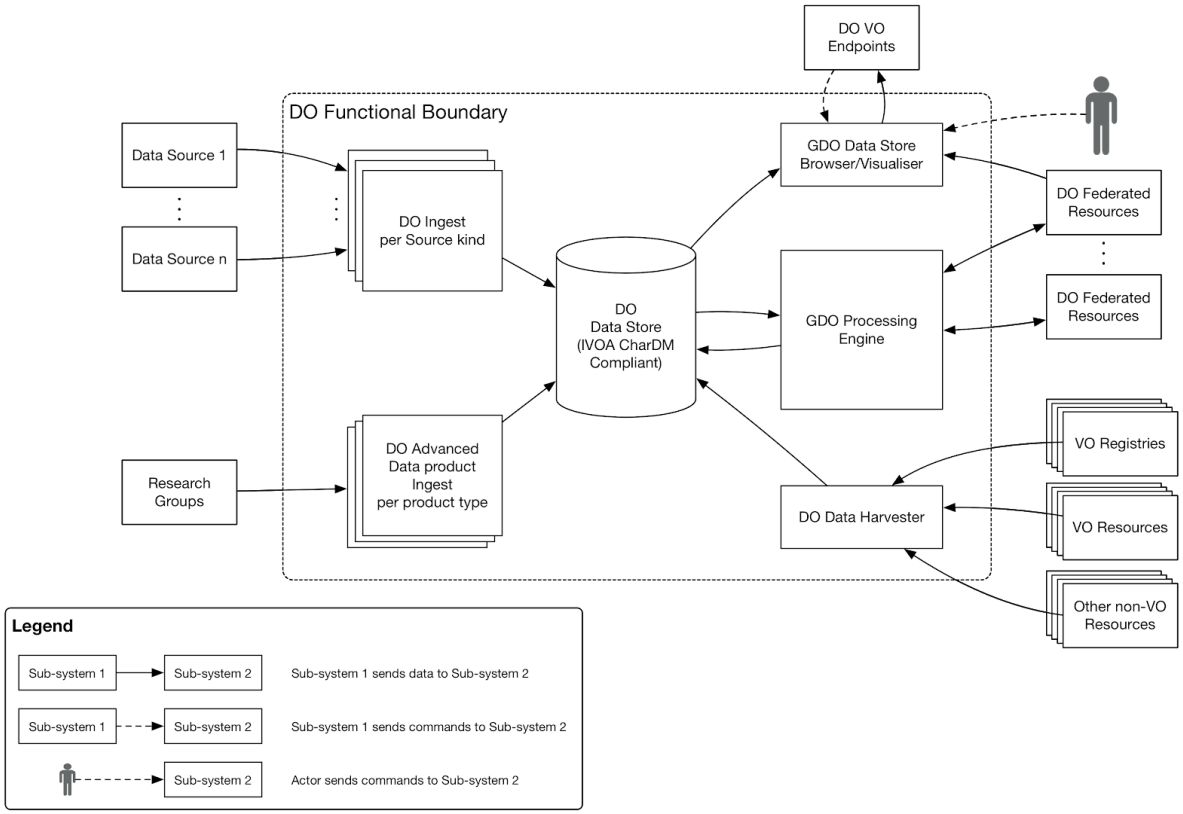

Figure 6: Functional View of the Data Observatory in the Astronomy use case.

Use Case: Chilean ALMA Centre

Data produced by the ALMA Observatory will be ingested to the DO after it goes through the ALMA pipeline in the Observatory operations. There will be a subset of *Data Ingestion tools* at the Tools layer, specially designed to ingest ALMA both the Science Data Models (ASDMs) through Oracle GateKeeper replication and the relevant bulk data through NGAS replication. Both Oracle GateKeeper and NGAS live

as containers in the Abstract Processing layer, while the Oracle database lives at the Datalake layer. Given the strict requirements from ALMA in terms of the processing stack —exact versions of software, database schemas, Oracle version, CASA versions, etc.—, those containers/virtual machines will already be pre-populated, and available to run in any of the clouds as part of the Abstract Processing layer. The Abstract Storage layer must select an optimal way to store the data in DO infrastructure (multi-cloud + federated computing resources), but keeping again the strict latency requirements between processing and storage.

It is a mandate from the Joint ALMA Observatory (JAO) that all ALMA Regional Centres must comply with the ALMA Release policy, and with standards for the integration of each ARC with the JAO Dataflow and operations.

Use Case: DO Data Exploration

The data produced by the International Observatories that have a membership with the DO requires ingest processes which are tuned per instrument kind, and further per instrument. The ingest process lives at the Tool level, and prepares the data for ingestion in the Datalake, using DO APIs living at the Access and Processing APIs layer. The result of the process is data being ingested in the logical Datalake, and databases of the data, spatial indexing, and metadata ingesting being populated. In particular, the ObsCore database will be filled in, and ElasticSearch will be used for the indexing of all the relevant metadata.

The Abstract Processing layer is used by the DO APIs to instantiate the relevant processing resources from the different cloud providers, taking into account both the processing and storage needs.

The Datalake and Filesystem will use resources managed by the Abstract Storage layer to create a redundant storage from which the filesystem APIs can operate, and database records which can be queried.

A Data-gazer, DO trainee, DOer or Researcher can use a Jupyter Notebook portal existing in the Tool layer to explore the ingested data. The user will have available a number of example notebooks, which run against a Jupyter kernel.

The Jupyter Notebook portal uses DO APIs to secure some processing resources from the Abstract Processing layer that will be run on each kernel invocation. The environment will be based on the Python 3.x Anaconda distribution, with potentially R and Julia environments available.

The environment will allow both anonymous and non-anonymous access to the DO Datalake through Python abstractions of the Datalake/Filesystem APIs. The environment will also be able to do queries through the Virtual Observatory standard protocols both to DO data and to external datasets for combination and exploration.

It should be possible to have access to instances with higher performance for registered users so that operations such as mosaicking, co-adding (natural, weighted, averaged, median-based…) are possible. Customized environments will also be possible for registered users.

The DO APIs/Python environment will be able to recognise natively FITS, ALMA Science Data Model (ASDM)/Measurement Set (MS), and HDF5 at a minimum, with World-Coordinate System (WCS) support in all of them, and also in the relevant indexing databases, so that multiple spatial and temporal queries are possible, both on the ObsCore and the metadata tables.

The NOAO Datalab[5] is considered as a precursor of this Use Case.

Use Case: Precovery of Solar-System Body in the Data Observatory

To enable this use case, apart from the steps described in the first three paragraphs of the Use Case on data exploration using Jupyter Notebooks, additional metadata is required to be generated as part of the ingestion process.

For data which are ingested by the DO, in order to enable this Use Case, a series of metadata regarding standardized calculation of exposure time, the field of view, and actual region in STC (Space-Time Coordinates) standard for each field need to be generated as part of the DO Ingest Tools. An additional process can use a system such as SkyBoT[6] to identify potential, already known Solar-System Bodies in each of the ingested datasets, generating a searchable database of Solar-System Bodies available in the DO datasets. All of those processes live at the Tools layer. Bear in mind that this does not enable the identification of candidate datasets for precovery of newly characterized bodies.

From the orbital elements of a newly found body, it is possible to generate interpolated RA/Dec pairs for the extent of time of the DO dataset, so that candidate data can be identified. This requires an additional DO Tool based on something like JPL Horizons[7], or the OrbFit[8] package.

[5] https://datalab.noao.edu/
[6] http://vo.imcce.fr/webservices/skybot/
[7] https://ssd.jpl.nasa.gov/horizons.cgi
[8] http://adams.dm.unipi.it/orbfit/

The European Southern Observatory used to provide a system similar to the precovery of known SSBs, but only on Hubble Space Telescope (HST) data, and is no longer available. The Canadian Astronomy Data Centre (CADC) has a system called SSOIS[9] (Solar System Object Image Search) which does support finding candidate images for precovery based on orbital elements and date ranges.

[9] http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/ssois/

Annex 3: KPIs per line of work

The following tables include examples of KPIs per line of work. These KPIs are in the process of being developed. Timelines for their fulfillment will be determined once the founding process of the DO is completed.

Dataset and Dataset products

This group of KPIs is related to the dataset and how it is exploited by the different stakeholders. It includes KPIs related to the data lake itself and how it is exploited for science and business.

- **Data Lake**

This group of KPIs is related to the general performance of the data lake.

KPI	Definition	Stakeholders	Mission
Astronomical Data	1.- Amount of astronomical data (TB or PB) ingested by the D.O. 2.- Share of Astronomical Data compared to the total data lake.	Infrastructure, Data Gazers, Data Providers, Researchers, Research Organizations	6, 7
General Data	1.- Amount of non-astronomical data (TB or PB) ingested by the D.O. 2.- Share of non-astronomical data compared to the total data lake.	Infrastructure, Data Gazers, Data Providers, Researchers, Research Organizations	6, 7
Data Lake Growth	Percentual Increase of the Data Lake	Infrastructure, Data Gazers	6, 7
Data Lake Access (i)	1.- Access of group (i) to the data lake. Groups could be defined by different countries, stakeholders, etc. 2.- Share of the access of the group (i) compared to the total. ● 1 ratio has to be Chile/international.	Infrastructure, DOers, Government	6, 7
Ratio Online Usage Versus Downloads	Amount of data downloaded against data processed in the cloud.	Infrastructure, DOers.	1, 7
System Availability	TBD different technical KPIs.	DOers, Data gazers, Data enthusiasts, DO Trainees	6, 7

- **Science**

These KPIs are related to the science conducted with the data lake, both from researchers using the data and DOers advancing data science.

KPI	Definition	Stakeholders	Mission

Papers Using the DO	1.- Papers using data extracted or analyzed in the DO per year. 2.- Ratio of papers using downloaded data versus papers using DO tools and analyzing in the Cloud.	Researchers, Research Institutions, Infrastructure, Data Providers.	4
Papers generated by the DO	Papers generated by DOers from developments in Data Science in the DO.	Researchers, Research Institutions, DOers, Data enthusiasts.	1, 4
Cross-pollination	A measure of field cross-pollination, i.e. when a paper published in a field's journal used DO data from another field domain.	Researchers, Research Institutions, Data Enthusiast.	4

- **Products**

This group of KPIs is related to the products generated by the DO.

KPI	Definition	Stakeholders	Mission
Software Packages (i)	The number of products of the category i produced by the DO.	DOers, Large Enterprises, SMEs, Data Gazers	1, 4, 7
Use of software package (free).	A measure of use of free DO tools	DOers, Data enthusiasts, researchers,	1, 2, 6
Use of software package (paid).	A measure of use of paid DO tools	DOers, Large Enterprises, SMEs, Universities	1, 2

Challenge Management

KPI	Definition	Stakeholders	Mission
Challenges Generation	1.- Number of DO Challenges. 2.- Number of external challenges. 3.- DO-External Challenges ratio.	DOers, Data Gazers, Large Enterprises, SMEs, Universities, Researchers, Research organizations.	1, 2, 4
Industry involvement	The share of challenges that involve industry.	Large Enterprises, SMEs, Government	2, 3, 4
Open Innovation	The share of Challenges solved via open innovation.	Large Enterprises, SMEs, Data gazers, domain enthusiasts.	1, 2, 3, 7
Matching	The average time to find a match for the challenge.	DOers, Large Enterprises, SMEs.	2, 4
The Success rate of the matches.	A measure of the ability of the DO to find people or organizations available to work in a challenge.	DOers, Large Enterprises, SMEs.	2, 4

Radical Innovation potential	A measure of the innovation potential of each challenge.	Large Enterprises, SMEs, Research Facilities, Universities.	1
Knowledge Generation	A measure of the knowledge generated from challenges	DOers	1

Talent Formation

These KPIs are related to the educational resources generated by the DO and its ability to contribute to talent formation in data science.

KPI	Definition	Stakeholders	Mission
Courses engagement	Attendance to DO courses.	DOers, Large Enterprises, SMEs, DO Trainees, Universities, High education entities.	5, 6
Educational Resources generation	The number of educational resources generated by the DO.	Universities, High education entities, DO Trainees, Schools	5, 6
Educational Resources Usage	Access or download of educational resources generated by the DO.	Universities, High education entities, DO Trainees, Schools	5, 6
MOOCs	1- The number of subscriptions. 2- The number of completed courses.	Large Enterprises, SMEs, Universities, High education entities, DO Trainees, schools.	5, 6
Certificates	The number of certifications.	DO Trainees, Large Enterprises, SMEs, Schools.	5, 6
Educational areas distribution	Percentage of material and courses for each data science area.	DOers, Schools, DO Trainee, Universities, High education entities.	6

Outreach

These KPIs are related to the engagement of the DO with its different stakeholders and with its diffusion activities.

- **Communication to the general public**

This group of KPIs is aimed to measure how the DO is communicating what is being generated to the general public.

KPI	Definition	Stakeholders	Mission
Social networks	TBD depending on the social networks selected.	Data gazers, data enthusiasts, DOers.	6, 7

Web Page	1.- Number of visitors. 2.- Number of new visitors.	DOers	6
General Public Tools	Usage of general public tools.	Data gazers, Data enthusiasts, DOers, Government	6
General Public Products	The number of products oriented to the general public.	Data gazers, Data enthusiasts, DOers, Government	6

- **Stakeholders Management**

These group of KPIs is related to the relationship of the DO with its different stakeholders.

KPI	**Definition**	**Stakeholders**	**Mission**
Agreements	Agreements signed with the different data providers.	Data providers. DOers.	7
Memberships	The number of members.	DOers	2, 3, 4
Open Data	The share of the data lake that is open data.	Data gazers, Domain enthusiasts.	5, 6
Stakeholders involvement	Measure that the main stakeholders are participating in the committees.	Large Enterprises, SMEs, Universities, Higher education entities, Research Centers	3

Annex 4: DO Stakeholders

Data-gazer: A person which has curiosity on multiple datasets, and wants to learn how they are organized, and potentially obtain insight from the datasets. They would use the DO to find datasets of interest and to perform various analysis and transformations on the data to learn about them.

Domain Enthusiast: A person which has curiosity on the content produced by the DO datasets. They would use the public-ready datasets for recreational and educational purposes.

Data Producer: An entity that produces data which is considered for ingestion in the DO. It will initially consist of Astronomical Observatories and other DO members.

DO Trainees: Persons that will use the DO data, generated content, and/or tools for training themselves. They are not necessarily enrolled in any other School or Higher Education Entity, or members of any kind of Enterprise.

DOers: People working for the DO that will work with the data lake, develop solutions and advise challenges.

High Education Entity: An organization whose main goal is to provide training beyond what is provided by the mandatory education, but regulated by the Ministry of Education. Universities and Technical Training Centres (Centros de Formación Técnica in Spanish) belong to this class of entities for the purpose of DO operation.

Infrastructure as a Service Provider: An entity that provides Storage and/or Processing services which are accessible from (almost) anywhere in the world with Internet connectivity.

Large Enterprise: The Ministry of Economy of Chile defines a Large Enterprise[10] as a company with sales larger than 100.000 Fiscal Units[11] (UF) per year.

Research Organization: An organization whose main goal is to foster the development of knowledge, irrespective of whether they generate profits or not. Research Organisations are not necessarily Higher Education Entities.

[10] http://www.sii.cl/contribuyentes/empresas_por_tamano/pymes.pdf
[11] The value of the UF is currently around 27600 Chilean pesos per UF on December 2018. The official, up-to-date values for the UF can be found at http://www.sii.cl/valores_y_fechas/uf/uf2018.htm.

Researcher: A person, belonging or not to a Research Organization, that can potentially use data from the DO to generate new knowledge, probably linked with a particular research field or proposal.

School: An organization whose main goal is to provide training within the mandatory education curriculum. Their typical use case for the DO is outreach, and using it for relevant training using astronomy or other disciplines within the

Small and Medium-sized Enterprise (SME): The Ministry of Economy of Chile defines a SME[12] (PyME in Spanish) as a company with sales between entre 2.400 and 100.000 Fiscal Units[13] (UF) per year.

University: A High Education Entity which is also a Research Organisation.

Government: Ministries that have interests in the DO like the Ministry of Economy, of Foreign Affairs, Of Science, Technology, Knowledge and Innovation, of Education, etc.

[12] http://www.sii.cl/contribuyentes/empresas_por_tamano/pymes.pdf
[13] Assuming around 27600 Chilean pesos per UF on December 2018, the sales threshold for an SME is between 6.6 and 2760 million pesos.

Annex 5: Glossary

Container: a way of packaging processing software that includes basic operating system services that can be run in a containerized environment, such as the processing instances of a Cloud Provider.

Datalake/Data Lake: Logical concept in which data from different provenances can be stored together, but tagged in such a way that differentiated processing can happen on arrival to the Datalake. Sources to a Datalake includes the results of processing in the Datalake itself.

Annex 6: List of Acronyms

ALMA	Atacama Large Millimetre and sub-millimetre Array
API	Application Program Interface
ASDM	ALMA Science Data Model
CADC	Canadian Astronomy Data Centre
DO	Data Observatory
ESO	European Southern Observatory
FITS	Flexible Image Transport System
HDF5	Hierarchical Data Format 5
HST	Hubble Space Telescope
IP	Intellectual Property (property rights context)
IP	Internet Protocol (networking context)
IP	Investigador Principal (astronomy observations context; Spanish for Principal Investigator)
IVOA	International Virtual Observatory Alliance
JAO	Joint ALMA Observatory
MS	Measurement Set
NOAO	National Optical Astronomy Observatory
NPO	Non-Profit Organization
NRAO	National Radio Astronomy Observatory
PI	Principal Investigator
PyME	Pequeña y Mediana Empresa (Spanish for SME)
SME	Small-to-Medium Enterprise
SSB	Solar System Body
TBC	To Be Confirmed

TBD	To Be Defined
TBW	To Be Written
UF	Unidad Fiscal (Fiscal Unit)
VO	Virtual Observatory

References

Cameron Partners Innovation Consultants, *Estudio de Nuevos negocios y spin-offs a partir de la astroingeniería,* 2013.

Catanzaro, M., Miranda, G., Palmer, L. & Bajak, A., *South American science: Big players*, in Nature (2014)

CONICYT. *Roadmap for the Fostering of Technology Development and Innovation in the Field of Astronomy in chile*, 2012.

De Montis, A., De Toro, P., Droste-Franke, B., Omann, I., & Stagl, S. *Criteria for quality assessment of MCDA methods*, In 3rd Biennial Conference of the European Society for Ecological Economics, Vienna (pp. 3-6), 2000.

EY, *Astroinformatics Business Opportunities (actual offer, future demand, and opportunities for Chile)*, commissioned by CORFO Astroinformatics program, February 2018.

Guridi, J. A., Pertuzé, J., Pfotenhauer, S., *Natural Laboratories as Policy Instruments for Technological Learning and Institutional Capacity Building: The Case of Chile's Astronomy Cluster*, submitted for publication to Research Policy, 2018.

Quinn, P., Axelrod, T., Bird, I., Dodson, R., Szalay, A., and Wicenec, A., *Delivering SKA Science*, Proceedings in Advancing Astrophysics with the Square Kilometre Array (AASKA14), 2015.

Schwab, Klaus, *The Fourth Industrial Revolution*, ISBN 978-1524758868, 2017.

Szalay, A. S., *The National Virtual Observatory*, in Astronomical Data Analysis Software and Systems X, 2001.