

Estado da publicação: O preprint não foi publicado em outro meio.

From startup registry to observatory: data quality in southern Brazil

Rafael Baptista

<https://doi.org/10.1590/SciELOPreprints.18203>

Submetido em: 2026-09-28

Postado em: 2026-09-30 (versão 1)
(AAAA-MM-DD)

From startup registry to observatory: data quality in southern Brazil

Rafael Reimann Baptista

Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil

ORCID: <https://orcid.org/0000-0003-1937-6393> · E-mail: rafael.baptista@pucrs.br

Preprint. Manuscript submitted to the Revista de Administração Contemporânea (RAC). Funding: CNPq Research Productivity Grant (DT).

Portuguese title: Do cadastro ao observatório: qualidade dos dados de startups no Rio Grande do Sul

ABSTRACT

Objective: to diagnose the quality of a public state registry of startups and to propose, and test on the same data, a longitudinal observatory able to monitor startup ecosystems over time.

Method: the study analysed all 1,514 records of the startup registry of Rio Grande do Sul, Brazil, extracted from its public report in September 2026. A protocol written before the analyses defined indicators for six dimensions of data quality: completeness, consistency, uniqueness, validity, coverage and timeliness. Candidate duplicates and the mapping of market segments were coded independently by two coders, and the registration dates of 693 website domains were retrieved from the Brazilian domain registry.

Results: half of the records had all descriptive fields filled, and business model was missing in 44.6%. Missingness varied sharply across municipalities, from none to 81.0% for business model, which points to differences between the organizations that register startups. The registry had no identifier, no dates and no record of the information source. Only 31.6% of the records would pass ten validation rules proposed for the observatory, and 31.5% could be found again in a future wave only by name.

Conclusions: the registry is a reasonable snapshot but cannot yet support longitudinal analysis. The proposed observatory links each diagnosed failure to a design component: persistent identifier, dated observations with source, controlled vocabularies, validation at entry and a six-monthly wave calendar. The procedure can be replicated by other regional programmes that maintain startup registries.

Keywords: data quality; startups; entrepreneurial ecosystems; innovation policy; administrative records.

RESUMO

Objetivo: diagnosticar a qualidade de um cadastro público estadual de startups e propor, testando sobre os mesmos dados, um observatório longitudinal capaz de acompanhar ecossistemas de startups ao longo do tempo.

Método: foram analisados os 1.514 registros do cadastro de startups do Rio Grande do Sul, extraídos do relatório público em setembro de 2026. Um protocolo escrito antes das análises definiu indicadores para seis dimensões de qualidade de dados: completude, consistência, unicidade, validade, cobertura e atualidade. Os candidatos a duplicidade e o mapeamento dos segmentos de mercado foram codificados de forma independente por dois codificadores, e as datas de registro de 693 domínios foram obtidas no registro brasileiro de domínios.

Resultados: metade dos registros tinha todos os campos descritivos preenchidos, e o modelo de negócio faltava em 44,6%. A ausência variou muito entre municípios, de zero a 81,0% no modelo de negócio, o que indica diferenças entre as organizações que cadastram as startups. O cadastro não tinha identificador, datas nem registro da fonte da informação. Apenas 31,6% dos registros passariam em dez regras de validação propostas para o observatório, e 31,5% só poderiam ser reencontrados numa próxima onda pelo nome.

Conclusões: o cadastro é uma fotografia razoável, mas ainda não permite análise longitudinal. O observatório proposto liga cada falha diagnosticada a um componente de desenho: identificador persistente, observações datadas com fonte, vocabulários controlados, validação na entrada e calendário semestral de ondas. O procedimento pode ser replicado por outros programas regionais que mantêm cadastros de startups.

Palavras-chave: qualidade de dados; startups; ecossistemas empreendedores; política de inovação; registros administrativos.

INTRODUCTION

Governments that support startups need to know how many startups exist, where they are, what they do and how they change over time. The entrepreneurial ecosystem approach made this need explicit. It treats the startup population as an outcome of interdependent conditions such as networks, finance, talent, knowledge and support services, and it asks policy to monitor those conditions and their results over time (Acs, Stam, Audretsch, & O'Connor, 2017; Stam, 2015; Stam & van de Ven, 2021). Monitoring, however, depends on data. The measurement literature has shown that ecosystem indicators are sensitive to how the underlying data are collected and aggregated, and that indicators built from convenient but

incomplete sources can mislead (Brown & Mason, 2017; Leendertse, Schrijvers, & Stam, 2022).

Most research on startups relies on commercial databases, such as Crunchbase, or on directories assembled by business associations. These sources are valuable, but their coverage depends on voluntary reporting and on the visibility of each company, which favours firms that are larger, older, funded or located in dense hubs (Dalle, den Besten, & Menon, 2017; Retterath & Braun, 2020). Regional governments in emerging economies increasingly build their own registries to fill this gap. In Brazil, where startups gained a legal definition with the 2021 legal framework for startups (Lei Complementar n. 182, 2021), state innovation programmes map the companies in their territories and publish the results in dashboards. These registries are rarely evaluated as data. They are consulted as lists, not audited as measurement instruments.

This article addresses that gap for the startup registry of Rio Grande do Sul, a state in southern Brazil whose innovation policy, the Inova RS programme, organizes the territory into eight regional innovation ecosystems (Decreto n. 54.767, 2019; Prochnow, Vieira, Baggio, & Thesing, 2025). The registry, maintained by the state network Rede RS Startup, lists more than 1,500 startups and is the only statewide source on the population of startups. It is also a single snapshot. It carries no identifier, no date and no record of who supplied each piece of information, so it cannot show which startups entered, grew, moved or disappeared. The problem that motivates this study is therefore practical: a public registry that could become the backbone of ecosystem monitoring is currently unable to support any comparison over time.

The research question is: what is the quality of the information in a public state registry of startups, and which design requirements does this diagnosis impose on a longitudinal observatory of startup ecosystems? The objective is to diagnose the quality of the Rio Grande do Sul registry in six dimensions and to propose, and test on the same data, the architecture of an observatory that corrects the failures found.

The study makes three contributions. First, it translates the general data quality framework (Batini, Cappiello, Francalanci, & Maurino, 2009; Wang & Strong, 1996) into a set of indicators that fit a startup registry and that any programme can compute with public information. Second, it shows that missing information in the registry is not a random nuisance. Missingness varies strongly across municipalities and with the declared stage of the startup, which makes the registry more visible for some parts of the ecosystem than for others.

Third, it offers an observatory design in which each component answers a diagnosed failure, together with a proof of concept that measures how far the current data are from the proposed standard. From a managerial point of view, the results give programme managers a concrete list of changes, with an estimate of their effect, before any new data collection is funded.

CONTEXT AND INVESTIGATED REALITY

The Inova RS programme and the state startup registry

Rio Grande do Sul had 10.9 million inhabitants in the 2022 census and has 497 municipalities. Its state innovation programme, Inova RS, was created in 2019 to coordinate universities, companies, government and civil society in eight regional innovation ecosystems, each with its own governance and priorities (Decreto n. 54.767, 2019; Prochnow et al., 2025). Within this programme, the state Secretariat of Innovation, Science and Technology coordinates Rede RS Startup, a network that maps the startups of the state and publishes the result as an interactive map in a public Power BI report (SICT-RS, 2026).

The report displays, for each startup, eight fields: name, development stage (ideation, validation, growth or scale), market segment, vertical (the type of solution, such as HealthTech or AgTech), target market (for example business to business or business to consumer), business model, municipality and a website or social media address. The report does not describe how records enter the registry, how often they are revised or who is responsible for each field. It is used by programme managers, support organizations and researchers as the reference list of startups in the state.

Data and extraction

The working file was generated on 2 September 2026 by automated navigation of the public report, reading every row of its data table. The extraction was repeated three times with different scrolling settings and always returned the same 1,514 records, from 101 municipalities. The summary card of the same report announced 1,649 startups. The reason for the difference of 135 records could not be determined from the public interface. The original file was preserved without changes, and every derived variable was created by numbered scripts.

Research design

The study follows the logic of design-oriented management research, in which a field problem is diagnosed, a solution is designed from the diagnosis and the solution is tested

against the problem it is meant to solve (Hevner, March, Park, & Ram, 2004; van Aken, 2004). Here the field problem is the inability of the registry to support longitudinal monitoring, the designed solution is an observatory, and the test is a proof of concept that applies the observatory rules to the current records.

An analysis protocol was written and dated on 27 September 2026, before the collection of external data, the coding of duplicates and the production of any final table. The protocol fixed the indicators, the rules for detecting candidate duplicates, the threshold used to flag possibly outdated stages and the statistical procedures. Operational adjustments made afterwards, none of which changed the indicators or tests, are listed in the Supplementary Material. The analysis is descriptive. During the feasibility check that preceded the protocol, the author examined cross-tabulations between missing fields and declared stage. For this reason, associations between data quality and stage are reported as descriptions and not as confirmatory tests.

DIAGNOSIS OF THE PROBLEM SITUATION

Data quality as fitness for use

Data quality is usually defined as fitness for use: data are of good quality when they serve the purpose of those who use them (Wang & Strong, 1996). This definition matters for startup registries because the same list can be adequate for one use, such as sending an invitation to all registered startups, and inadequate for another, such as estimating how many startups reached the growth stage in the last two years. The literature groups the attributes of data quality into dimensions. Completeness refers to the presence of the values needed for a task; consistency to the agreement of values with rules and with each other; uniqueness to the absence of duplicated entities; validity or accuracy to the correspondence between the recorded value and the real object; and timeliness to how current the values are (Batini et al., 2009; Pipino, Lee, & Wang, 2002). Coverage, the relation between the registered population and the target population, is added when the data are used to describe a population rather than to manage individual cases (Groves & Lyberg, 2010).

Registries built by public programmes resemble administrative records more than surveys. They are collected for operational purposes, their contents depend on the procedures of the organizations that feed them, and they are later reused for statistics (Hand, 2018; Zhang, 2012). Two consequences follow. First, the errors in such data are often structured by the collection process, not by the units described, so that missing values may reveal who

collected the data rather than what the startup is like. Missing values that depend on observed or unobserved characteristics are not ignorable in analysis (Little & Rubin, 2019). Second, reusing administrative data over time requires stable identifiers and dated observations, which operational lists often lack. Business demography statistics, for instance, depend on identifying the same enterprise across years to separate births, deaths and continuing units (Eurostat & OECD, 2008), and the linkage of records without a common key is itself an error-prone step (Christen, 2012; Fellegi & Sunter, 1969).

For startups, the web has become a complementary source. Company websites have been used to identify innovative firms, map innovation ecosystems and track entry into emerging technologies (Arora, Youtie, Shapira, Gao, & Ma, 2013; Axenbeck & Breithaupt, 2021; Gök, Waterworth, & Shapira, 2015; Kinne & Axenbeck, 2020). Web data are cheap and current, but they are also selective, since firms differ in how much they show online (Blazquez & Domenech, 2018). An online presence can work as a signal of quality to customers and investors (Connelly, Certo, Ireland, & Reutzel, 2011), which means that the information available about a startup may itself vary with its maturity. The diagnosis below combines these perspectives: it treats the registry as administrative data, uses the web address as both a field to be evaluated and a source of external dates, and asks whether the gaps are concentrated in specific parts of the ecosystem.

Diagnostic procedures

All 1,514 records were analysed, including those without a declared stage. Table 1 lists the indicators by dimension. Completeness was measured by indicators of absence for the six informative fields and by a count of filled descriptive fields. Consistency was assessed in three ways: the number of labels per field and the presence of synonyms or of the same category written in two languages; and the coherence between vertical and segment, defined as both pointing to the same sectoral family in a classification of nine families (health and life sciences; agriculture and food; information technology; education; financial services; management and marketing; industry, construction and mobility; energy, environment and impact; consumer and experiences). Records whose vertical belonged to the information technology family, or whose segment named only a technology or a generic service, were excluded from the coherence denominator, because a technological vertical can legitimately serve any segment.

Uniqueness was assessed with three rules fixed in the protocol: identical names after normalization, the same website domain or intermediary page, and a Jaro-Winkler similarity

of 0.92 or more between names in the same municipality (Winkler, 1990). Candidate pairs were classified into four categories: the same startup registered more than once; different products or units of one company; a shared address belonging to an intermediary such as an incubator; and distinct entities (Elmagarmid, Ipeirotis, & Verykios, 2007). The mapping of the 54 segments onto the sectoral families required the same kind of judgement. Both tasks were coded by the author with the assistance of a large language model (Claude, Anthropic) and recoded independently, blind to the first coding, by a model from a different developer (ChatGPT, OpenAI) in a temporary session. Agreement was measured with Cohen's (1960) kappa, with bootstrap confidence intervals, and the author decided all disagreements. Language models have shown good agreement with human coders in annotation tasks, but their use as coders requires transparency and human adjudication (Gilardi, Alizadeh, & Kubli, 2023; Ziems et al., 2024).

Validity covered the type of web address (own domain, social media profile, other platform, intermediary page), the syntax of the address and signs that a legal name, rather than a trade name, had been recorded. For every own domain ending in .br, the registration date, the date of the last change, the expiry date and the status of the domain were retrieved from the public registration data access protocol (RDAP) service of the Brazilian domain registry on 27 September 2026, between 19:06 and 19:29 UTC. The 693 registrable domains were queried one at a time, and no query failed. Coverage was assessed against the total announced by the report and against the 497 municipalities of the state. Timeliness was assessed by the presence of dates in the registry and, indirectly, by the age of the domains. As a descriptive plausibility check fixed in the protocol, the age of the domain was compared across declared stages, and startups declared in ideation or validation with a domain registered more than eight years before the extraction were flagged as possibly outdated.

Finally, to examine whether quality depended on the collection channel, each indicator was compared across the eight regional ecosystems and across the 17 municipalities with at least 20 records, with chi-square tests whose p values were obtained by Monte Carlo simulation with 10,000 replicates, Cramér's V as effect size and the Benjamini and Hochberg (1995) correction within each family of tests. Proportions are reported with 95% Wilson (1927) confidence intervals. The analyses were run in Python 3.10 with pandas, NumPy, SciPy and statsmodels, with a fixed random seed.

Results of the diagnosis

Table 1 summarizes the quality profile of the registry, and Figure 1 displays the same indicators with their confidence intervals.

Table 1

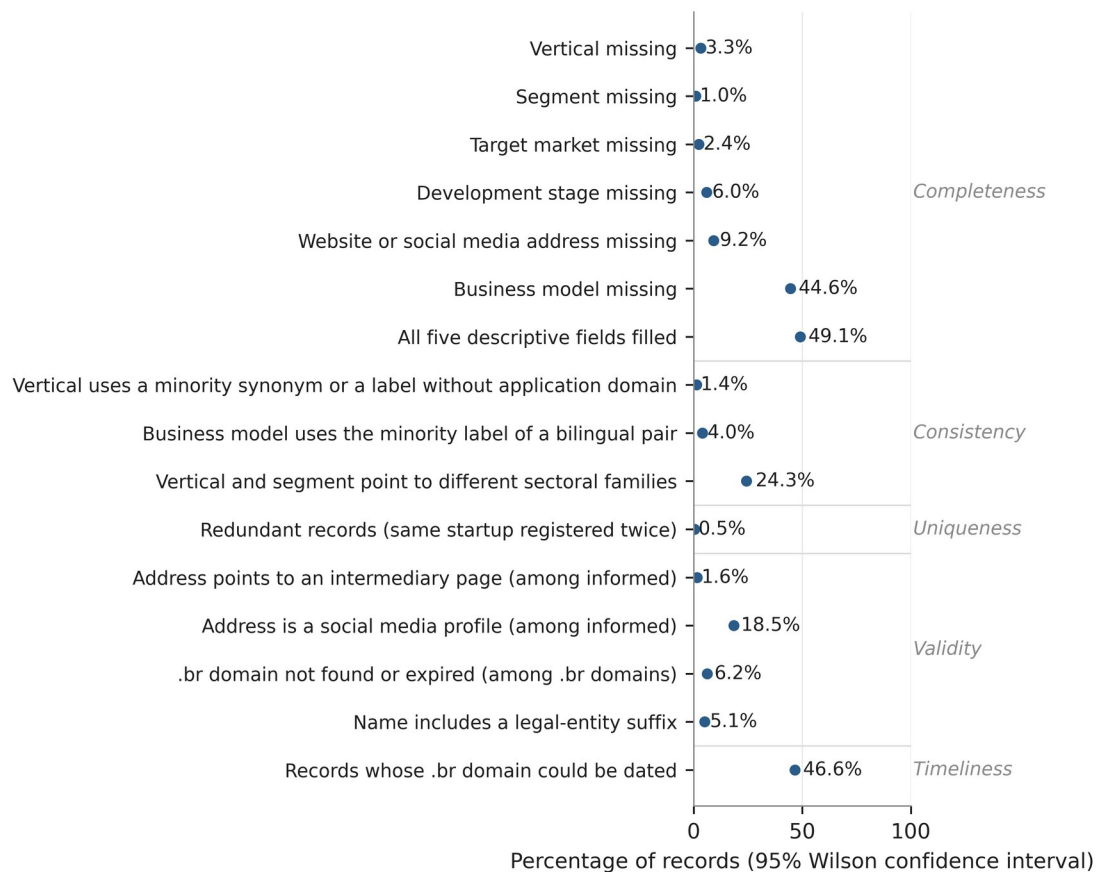
Quality profile of the startup registry of Rio Grande do Sul, by dimension.

Dimension	Indicator	n / N	% (95% CI)
Completeness	Vertical missing	50 / 1,514	3.3 (2.5 to 4.3)
	Segment missing	15 / 1,514	1.0 (0.6 to 1.6)
	Target market missing	36 / 1,514	2.4 (1.7 to 3.3)
	Development stage missing	91 / 1,514	6.0 (4.9 to 7.3)
	Website or social media address missing	140 / 1,514	9.2 (7.9 to 10.8)
	Business model missing	675 / 1,514	44.6 (42.1 to 47.1)
	All five descriptive fields filled	743 / 1,514	49.1 (46.6 to 51.6)
	Vertical uses a minority synonym or a label without application domain	21 / 1,514	1.4 (0.9 to 2.1)
Consistency	Business model uses the minority label of a bilingual pair	60 / 1,514	4.0 (3.1 to 5.1)
	Vertical and segment point to different sectoral families	210 / 864	24.3 (21.6 to 27.3)
	Redundant records (same startup registered twice)	8 / 1,514	0.5 (0.3 to 1.0)
Uniqueness	Address points to an intermediary page (among informed)	22 / 1,374	1.6 (1.1 to 2.4)
	Address is a social media profile (among informed)	254 / 1,374	18.5 (16.5 to 20.6)
Validity	.br domain not found or expired (among .br domains)	43 / 693	6.2 (4.6 to 8.3)
	Name includes a legal-entity suffix	77 / 1,514	5.1 (4.1 to 6.3)
	Records listed out of the total announced (1,649)	1,514 / 1,649	91.8
Coverage	Records with an observation or update date	0 / 1,514	0.0 (0.0 to 0.3)
	Records whose .br domain could be dated	705 / 1,514	46.6 (44.1 to 49.1)

Note. N = 1,514 records unless stated otherwise. Coherence between vertical and segment was assessed among 864 eligible records. Validity of .br domains refers to 693 registrable domains queried in the Brazilian domain registry on 27 September 2026. Wilson confidence intervals. Source: prepared by the author from the Rede RS Startup registry (extracted on 2 September 2026).

Figure 1

Quality profile of the startup registry of Rio Grande do Sul.



Dots are percentages of records and bars are 95% Wilson confidence intervals; denominators as in Table 1.
Source: prepared by the author from the Rede RS Startup registry and the Brazilian domain registry.

Completeness was high for the classification fields and low for the business model. Vertical, segment and target market were missing in 3.3%, 1.0% and 2.4% of the records, the development stage in 6.0% and the web address in 9.2%. The business model was missing in 44.6% (95% CI 42.1 to 47.1). Only 743 records (49.1%) had all five descriptive fields filled, and 9 records had nothing beyond name and municipality. The absences of vertical, segment and target market tended to occur together (ϕ between 0.31 and 0.51), whereas the absence of business model was almost unrelated to the others (ϕ at most 0.14). This pattern suggests two distinct mechanisms: a small group of records that entered the registry with minimal information, and a single field that is filled or skipped depending on the procedure followed.

Consistency problems were modest in frequency but revealing in form. The registry used 56 vertical labels, 54 segment labels, 24 business model labels and 8 target market labels. Several vertical labels were synonyms (for example LawTech and LegalTech, FinTech and Techfin), one label did not describe any application domain, and the same business model appeared in Portuguese and in English (Transação and "Transational"; Subscription and Economia de Recorrência). The code for missing information was written in two ways across

fields. Records using the minority label of a synonym pair were few (1.4% for vertical and 4.0% for business model), but their presence suggests that the categories were entered as free text or from lists that changed over time, rather than from a controlled vocabulary. Among the 864 records eligible for the coherence check, vertical and segment pointed to the same sectoral family in 75.7% (95% CI 72.7 to 78.4). Coherence ranged from 88.4% among health startups to 64.7% among management and marketing startups, which reflects both classification errors and the ambiguity of some categories.

Uniqueness was the strongest dimension. The three detection rules produced 27 groups with 61 records and 51 candidate pairs. Both coders agreed on all 51 pairs ($\kappa = 1.00$). Twenty-seven pairs shared the page of an incubator, science park or university, eleven were different products or units of the same company, eight were the same startup registered twice, and five were distinct companies with a coincident name or an address recorded by mistake. The registry therefore contained eight redundant records (0.5%). For the mapping of segments, the coders agreed on 50 of 54 segments ($\kappa = 0.91$; 95% CI 0.82 to 0.98).

Validity revealed how the web address mixes different kinds of information. Among the 1,374 records with an address, 77.9% pointed to an own domain, 18.5% to a profile on Instagram, LinkedIn or Facebook, 2.0% to other platforms and 1.6% to the page of an intermediary. The intermediary pages are an indirect trace of the collection channel, which the registry itself does not record. Of the 693 .br domains queried, 28 did not exist in the domain registry and 15 had expired without renewal between June and August 2026. Together, 6.2% of the .br domains (95% CI 4.6 to 8.3) no longer identified an active website, a proportion that sets a lower bound for the obsolescence of addresses, since a domain can remain active after a company stops operating. Legal-entity suffixes appeared in 5.1% of the names and 21.9% of the names were written entirely in capitals, both signs that legal and trade names were mixed.

Coverage could only be assessed internally. The table contained 91.8% of the total announced by the same report, and registered startups were present in 101 of the 497 municipalities (20.3%). The registry is not a census, and the absence of a documented target population prevents any estimate of undercoverage.

Timeliness was the weakest dimension, because the registry holds no date of registration, update or change of stage. The domain registry provided a partial substitute for 705 records (46.6%) whose own .br domain could be dated. The median age of these domains at the date of extraction was 6.8 years (interquartile range 4.4 to 10.2), and the median time since the last change was 7.7 months. Domains of startups declared in scale were older

(median 8.0 years) than those in the other stages (6.4 to 6.6 years; Kruskal-Wallis $H = 11.04$, $p = 0.012$; Spearman $\rho = 0.13$). Among the 239 startups declared in ideation or validation with a dated domain, 83 (34.7%; 95% CI 29.0 to 41.0) had a domain registered more than eight years earlier. This does not prove that their stage is outdated, because a domain may predate the startup when a company pivots or reuses an address. It shows, however, that a declared stage without a date cannot be interpreted as a current state.

Heterogeneity across collection channels

Table 2 and Figure 2 show that quality was not uniform across the territory.

Table 2

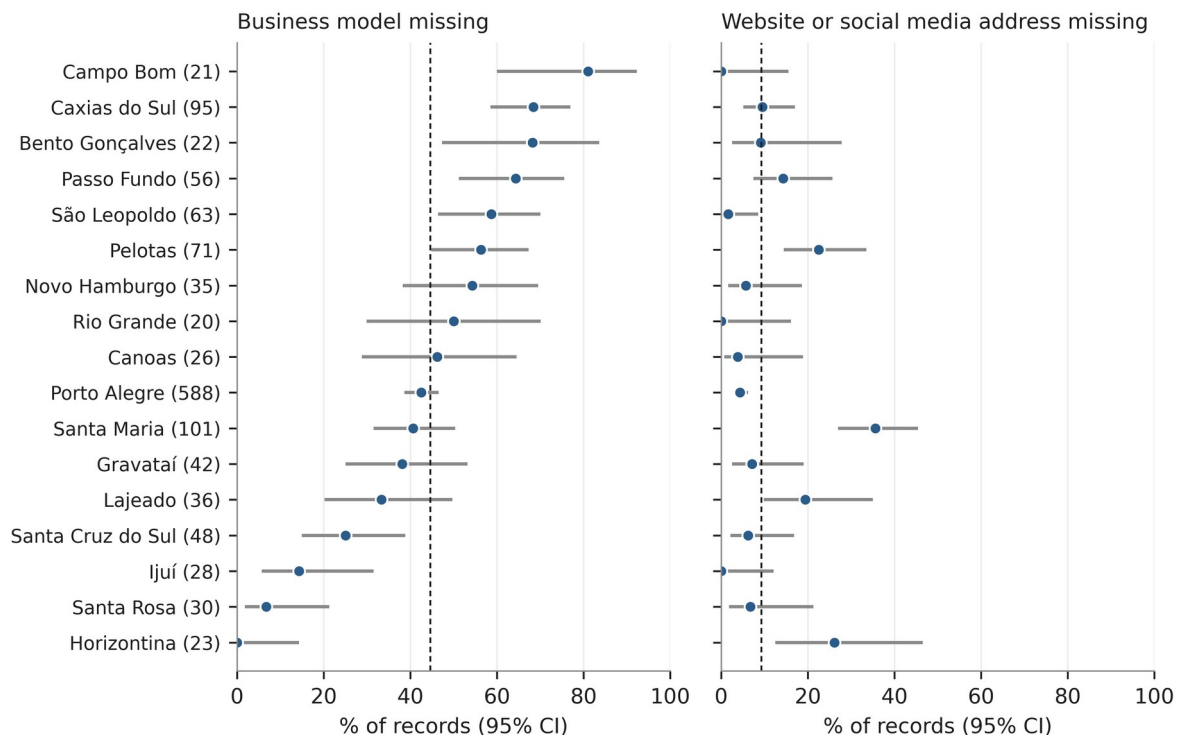
Variation of quality indicators across regional innovation ecosystems and municipalities.

Indicator	Regions (8; n = 1,514): range, %	V	Municipalities with 20 or more records (17; n = 1,305): range, %	V
Stage missing	0.0 to 11.9	0.13	0.0 to 15.8	0.16
Vertical missing	0.0 to 6.8	0.09 (ns)	0.0 to 7.9	0.11 (ns)
Segment missing	0.0 to 4.6	0.16	0.0 to 6.3	0.20
Target market missing	0.0 to 8.6	0.17	0.0 to 10.0	0.20
Business model missing	10.6 to 64.2	0.25	0.0 to 81.0	0.30
Address missing	4.4 to 31.4	0.26	0.0 to 35.6	0.33
Own domain as address	48.4 to 79.6	0.27	41.7 to 82.9	0.30

Note. Range: lowest and highest percentage among groups. V: Cramér's V. Chi-square tests with Monte Carlo p values (10,000 replicates), corrected by the Benjamini-Hochberg procedure within each family; all tests had adjusted $p < 0.05$ except those marked ns. Source: prepared by the author.

Figure 2

Missing business model and missing web address in the 17 municipalities with 20 or more records (number of records in parentheses).



Dots are percentages and bars are 95% Wilson confidence intervals; the dashed line marks the state average.
Source: prepared by the author.

The absence of business model ranged from 10.6% in the Noroeste e Missões region to 64.2% in the Serra Gaúcha region ($V = 0.25$), and from none of the 23 records of Horizontina to 81.0% of the 21 records of Campo Bom ($V = 0.30$). The absence of a web address ranged from 4.4% in the metropolitan region to 31.4% in the Central region, and from zero in several municipalities to 35.6% in Santa Maria ($V = 0.33$). All these differences remained after correction for multiple testing (adjusted $p < 0.001$) and were larger, not smaller, when Porto Alegre, the state capital with 588 records, was excluded ($V = 0.40$ for business model and 0.34 for address). The absence of vertical, in contrast, did not differ across regions or municipalities.

The municipal profiles also differed in which field was left empty. Horizontina had no missing business model but lacked the address in 26.1% of its records, whereas Campo Bom lacked the business model in 81.0% and the address in none. Differences of this size, in fields that are equally easy for any startup to report, are hard to attribute to the startups themselves. They are consistent with local collection procedures, in which each support organization fills the form in its own way. The registry thus measures, in part, the organizations that feed it.

Visibility also varied with the declared stage. The address was missing for 33.8% of the startups in ideation and for 2.3% of those in scale, and an own domain was recorded for

38.9% and 78.7% of them, respectively (Supplementary Table S1). The registry is therefore more informative about mature startups than about early ones, which is precisely the group whose trajectory an observatory would most want to follow.

Synthesis of the diagnosis

The registry is acceptable as a list: few duplicates, valid municipalities, classification fields almost always filled. It fails as a monitoring instrument for four reasons. It lacks the structural fields that make longitudinal use possible, namely an identifier, dates and the source of each record. Its missing values depend on the collection channel. Its vocabularies are not controlled. And the information it holds decays silently, as shown by the domains that no longer exist and by stages that may no longer be current. The next section proposes an observatory designed to correct each of these failures.

ANALYSIS AND PROPOSAL: A LONGITUDINAL OBSERVATORY OF STARTUP ECOSYSTEMS

Alternatives considered

Three alternatives to a redesigned registry were considered. The first is to rely on commercial or associative databases. They offer longer histories, but their coverage favours funded and visible companies and they are not governed by the public programme (Dalle et al., 2017; Retterath & Braun, 2020); they can complement, not replace, a public registry. The second is to monitor startups only through web data. Web mining is cheap and scalable (Kinne & Axenbeck, 2020), yet the diagnosis showed that online visibility is itself unequal across stages and territories, so a web-only system would reproduce the bias of the current registry. The third is a periodic survey of registered startups. Surveys can collect rich variables, but without a stable frame and identifier each survey would start from an unlinked list. The option retained here is to transform the registry itself into a versioned data system, using web and administrative sources as validation layers. This choice keeps the programme in control of its data and builds on an instrument that already has institutional support.

Design principles

The design follows four principles drawn from the literature on data governance and from the diagnosis. Data are treated as an asset with explicit owners, rules and quality targets (Abraham, Schneider, & vom Brocke, 2019; Khatri & Brown, 2010). Each observation carries its own date and source, as in register-based statistics, so that change can be separated from correction (Zhang, 2012). Categories are controlled and versioned, so that a change of label is

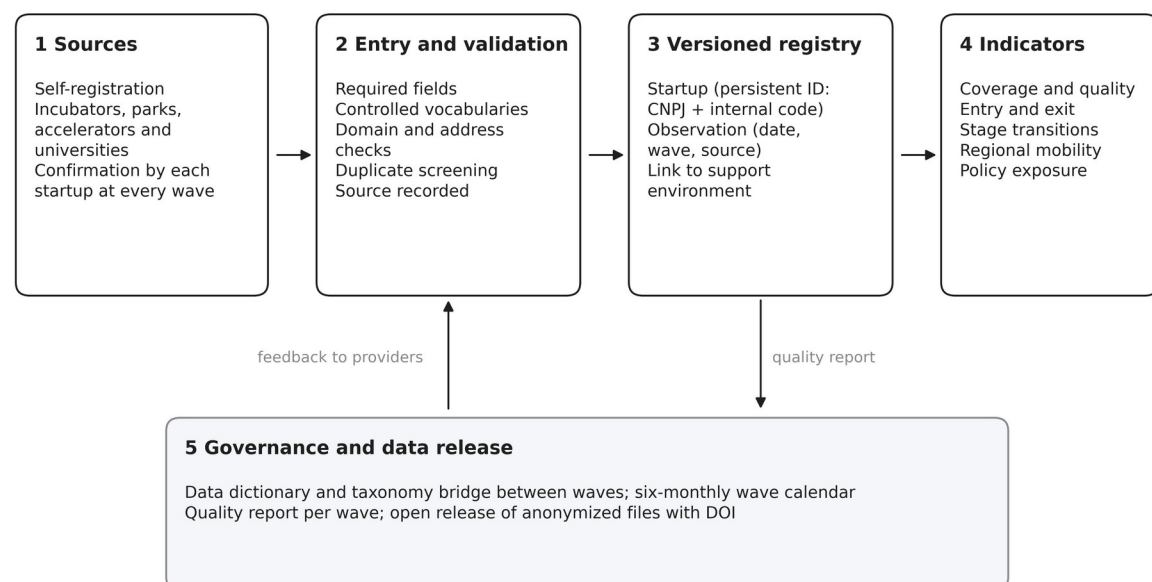
not read as a change of the economy, a problem well known in the classification of new sectors (Bertucci Ramos & Pedrosa, 2021; OECD & Eurostat, 2018). Released data follow the FAIR principles, being findable, accessible, interoperable and reusable (Wilkinson et al., 2016).

Components of the observatory

Figure 3 presents the architecture, and Table 3 links each diagnosed failure to the component that addresses it, the validation rule that implements it and the share of current records that would already comply.

Figure 3

Architecture of the proposed longitudinal observatory of startup ecosystems.



Numbers identify the five components described in the text. Source: prepared by the author.

Table 3

Diagnosed failures, components of the proposed observatory and compliance of current records (proof of concept).

Diagnosed failure	Observatory component	Rule or indicator	Current records complying, %
No identifier	Persistent identifier (CNPJ or internal code)	Record can be matched by a unique name and a unique own domain	68.1
No date or source for any value	Observation entity with date, wave and provider	Date and source recorded	0.0
Missing values that depend on the provider	Required fields and quality report returned to each provider	V01 Stage informed	94.0
		V02 Vertical informed	96.7
		V03 Segment informed	99.0

		V04 Target market informed	97.6
		V05 Business model informed	55.4
		V06 Address informed	90.8
Uncontrolled vocabularies	Controlled, versioned vocabularies with a bridge between waves	V07 Labels from the controlled vocabulary	91.3
Invalid or decaying addresses	Domain check at entry and at every wave	V08 Own domain, existing and active	67.8
Duplicated records	Duplicate screening at entry	V09 Not a redundant record	99.5
Legal and trade names mixed	Separate fields for legal and trade name	V10 Trade name without legal suffix	94.9
		All ten rules (V01 to V10)	31.6 (29.3 to 34.0)
		All rules except V05	56.1 (53.6 to 58.6)
Undocumented coverage	Definition of the target population and comparison with external sources	Records listed out of the total announced	91.8

Note. N = 1,514 records. Rules V01 to V10 were written as the entry specification of the observatory and applied to the current registry. Source: prepared by the author.

The first component defines the sources. Startups can register themselves or be registered by incubators, science parks, accelerators and universities, as today, but each record carries the identity of its provider, and each startup is asked to confirm or update its data at every wave. The confirmation turns the registry from a list that only grows into a panel in which absence of confirmation becomes information about activity.

The second component is entry and validation. Required fields, controlled vocabularies, syntax and domain checks, and duplicate screening are applied at the moment of entry, not after publication. The rules tested below were written as the specification of this component. Because the diagnosis showed that missingness depends on the provider, the validation reports are returned to each provider, which creates a feedback loop between data use and data collection.

The third component is the versioned registry, with three entities. The startup entity holds the persistent identifier, formed by the national company registration number (CNPJ) when it exists and by an internal code otherwise, since early startups may not yet be incorporated. The observation entity stores the values declared at each wave, with date and source. The link entity records the relation between the startup and support environments, such as incubators and parks, over time. This structure is what allows the indicators of business demography, such as entry, exit and survival, to be computed (Eurostat & OECD, 2008).

The fourth component produces indicators at state and regional level: coverage and quality, entry and exit, transitions between stages, mobility between municipalities and exposure to policy instruments. The fifth component is governance and release: a data dictionary, a bridge between taxonomies of successive waves, a six-monthly wave calendar, a quality report per wave and the open release of files without personal data, with a persistent identifier for each release.

Proof of concept

Ten validation rules were applied to the 1,514 current records to measure the distance between the registry and the proposed standard (Table 3). Compliance was above 90% for most individual rules: stage informed (94.0%), vertical informed (96.7%), segment informed (99.0%), target market informed (97.6%), address informed (90.8%), controlled vocabulary (91.3%), absence of duplication (99.5%) and trade name without legal suffix (94.9%). Two rules concentrated the failures: business model informed (55.4%) and a valid own domain as address (67.8%). Only 479 records (31.6%; 95% CI 29.3 to 34.0) passed all ten rules, and 850 (56.1%) passed all rules except business model. In other words, more than two in three records would need some correction before entering the observatory, and most of the corrections concern two fields.

The proof of concept also estimated linkability, the capacity to find the same startup in a future wave. Without an identifier, a record can only be matched by name or by domain. Of the current records, 68.1% had both a unique normalized name and a unique own domain, 31.5% could be matched only by name and 0.3% by neither. Names change and are written in different ways, as the duplicates showed, so matching by name alone requires probabilistic linkage and clerical review (Christen, 2012; Fellegi & Sunter, 1969). Nearly a third of the registry would thus enter a second wave with a fragile key. Adding the CNPJ or an internal identifier at the next confirmation would remove this weakness at a marginal cost.

Expected gains and implementation path

The expected gains are of three kinds. For monitoring, dated observations and a persistent identifier allow the programme to measure what it currently cannot: how many startups enter and leave, how long they remain in each stage and whether regional ecosystems differ in these rates. For management, provider-level quality reports show which organizations need support to improve their records, and the municipal heterogeneity found in this study indicates that such support can be targeted rather than spread across the whole

network. For research, a versioned and documented release would allow the kind of panel analyses that the entrepreneurial ecosystem literature calls for (Leendertse et al., 2022; Stam & van de Ven, 2021).

The implementation can be gradual. The first wave would add the identifier, the date and the source to the existing records and ask all startups to confirm their data, using the current registry as the frame. The rules in Table 3 can be implemented in the existing form at low cost. A six-monthly calendar yields five observations of each startup in two years, which is enough to estimate transition matrices between stages and to compare them across regions. The main costs are the maintenance of the form and the follow-up of non-respondents, and the main risk is attrition, which the observatory must itself measure by comparing startups that confirm with those that do not.

CONCLUSIONS AND TECHNOLOGICAL-SOCIAL CONTRIBUTION

This study asked how good the information in a public state registry of startups is and what an observatory built on it would require. The registry of Rio Grande do Sul is a reasonable list. Its classification fields are almost always filled, duplicates are rare and municipalities are valid. As a monitoring instrument, however, it has structural gaps: it has no identifier, no dates and no source, its missing values follow the collection channel, its vocabularies are not controlled and part of its information has silently aged. None of these problems is visible to a user who consults the public map, and all of them would bias an analysis that treated the registry as a census or as a panel.

The technological contribution of the study is a replicable procedure and a design. The procedure translates six dimensions of data quality into indicators that can be computed from public information, including external checks with the national domain registry, and it can be repeated at each wave as a quality report. The design links each diagnosed failure to a component of an observatory and specifies ten validation rules whose current compliance, 31.6% for all rules together, gives managers a baseline against which to measure improvement. The social contribution lies in the use of public data: a registry that knows who reported what and when can show whether support policies reach early-stage startups outside the main hubs, which the current snapshot cannot do.

For programme managers, three actions follow directly. The first is to add a persistent identifier, a date and a source to every record at the next update, since these three fields determine whether any longitudinal use will ever be possible. The second is to adopt

controlled vocabularies and entry validation for the fields that most often fail, business model and web address. The third is to return quality reports to the organizations that register startups, because the diagnosis indicates that quality depends on them. For researchers who use startup registries, the results recommend treating missing information as a property of the collection process, reporting it by territory and stage, and avoiding interpretations of declared stage as a current state when no date is available.

The study has limitations. It evaluates one registry at one point in time, so the proposed observatory was tested only on current data and its gains remain expected rather than observed. The extraction reproduced the public table, which contained fewer records than the total announced by the report, and the reason for the difference could not be established. The external checks covered only .br domains, which excludes startups represented by social media profiles or foreign domains, and the age of a domain is not the age of a company. The coding of duplicates and segments relied on a second coder that was a language model; agreement was high, but independence between models is limited and no second human expert was involved. Finally, the associations between data quality and declared stage are descriptive, since they were seen during the feasibility check.

Future work should implement the first wave of the observatory together with the organizations of Rede RS Startup, measure attrition and the effect of the validation rules on data quality, and compare the registry with other sources, such as company registration data, to estimate coverage. The same procedure can be applied to the registries that other Brazilian states and other countries maintain, making it possible to compare not only startup ecosystems but also the quality of the information on which they are judged.

DECLARATIONS

Data availability statement

Research data availability: the research data supporting the results of this study are available in the Zenodo repository at <https://doi.org/10.5281/zenodo.23015213>. The deposit contains the derived dataset (without addresses of personal profiles), the domain registry responses, the numbered analysis scripts, the analysis protocol and the coding files. The original registry is publicly available in the Rede RS Startup report.

Use of artificial intelligence: the author used Claude (Anthropic) to assist in writing the data processing scripts, as first coder of candidate duplicates and segments, and in drafting and editing the text, and ChatGPT (OpenAI) as an independent blind second coder. All analytical

decisions, adjudications and interpretations were made and checked by the author, who takes full responsibility for the content. Artificial intelligence tools are not listed as authors.

Conflict of interest: none.

REFERENCES

- Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 49, 424–438. <https://doi.org/10.1016/j.ijinfomgt.2019.07.008>
- Acs, Z. J., Stam, E., Audretsch, D. B., & O'Connor, A. (2017). The lineages of the entrepreneurial ecosystem approach. *Small Business Economics*, 49(1), 1–10. <https://doi.org/10.1007/s11187-017-9864-8>
- Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics*, 95(3), 1189–1207. <https://doi.org/10.1007/s11192-013-0950-7>
- Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites: Which website characteristics predict firm-level innovation activity? *PLOS ONE*, 16(4), e0249583. <https://doi.org/10.1371/journal.pone.0249583>
- Baptista, R. R. (2026). Replication package: Data quality of the Rio Grande do Sul startup registry and design of a longitudinal observatory [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.23015213>
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 1–52. <https://doi.org/10.1145/1541880.1541883>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bertucci Ramos, P. H., & Pedroso, M. C. (2021). Classification and categorization of Brazilian agricultural startups (Agtechs). *Innovation & Management Review*, 18(3), 237–257. <https://doi.org/10.1108/INMR-12-2019-0160>

- Blazquez, D., & Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130, 99–113.
<https://doi.org/10.1016/j.techfore.2017.07.027>
- Brown, R., & Mason, C. (2017). Looking inside the spiky bits: A critical review and conceptualisation of entrepreneurial ecosystems. *Small Business Economics*, 49(1), 11–30. <https://doi.org/10.1007/s11187-017-9865-7>
- Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer. <https://doi.org/10.1007/978-3-642-31164-2>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
<https://doi.org/10.1177/001316446002000104>
- Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39–67.
<https://doi.org/10.1177/0149206310388419>
- Dalle, J.-M., den Besten, M., & Menon, C. (2017). Using Crunchbase for economic and managerial research (OECD Science, Technology and Industry Working Papers, No. 2017/08). OECD Publishing. <https://doi.org/10.1787/6c418d60-en>
- Decreto n. 54.767, de 22 de agosto de 2019. (2019). Institui o Programa Inova RS. Governo do Estado do Rio Grande do Sul.
- Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 1–16.
<https://doi.org/10.1109/TKDE.2007.250581>
- Eurostat, & OECD. (2008). Eurostat-OECD manual on business demography statistics. OECD Publishing. <https://doi.org/10.1787/9789264041882-en>
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210.
<https://doi.org/10.1080/01621459.1969.10501049>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>

- Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, 102(1), 653–671. <https://doi.org/10.1007/s11192-014-1434-0>
- Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879. <https://doi.org/10.1093/poq/nfq065>
- Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), 555–605. <https://doi.org/10.1111/rssa.12315>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1), 148–152. <https://doi.org/10.1145/1629175.1629210>
- Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, 125(3), 2011–2041. <https://doi.org/10.1007/s11192-020-03726-9>
- Leendertse, J., Schrijvers, M., & Stam, E. (2022). Measure twice, cut once: Entrepreneurial ecosystem metrics. *Research Policy*, 51(9), 104336. <https://doi.org/10.1016/j.respol.2021.104336>
- Lei Complementar n. 182, de 1º de junho de 2021. (2021). Institui o marco legal das startups e do empreendedorismo inovador. Presidência da República.
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. <https://doi.org/10.1002/9781119482260>
- OECD, & Eurostat. (2018). *Oslo manual 2018: Guidelines for collecting, reporting and using data on innovation* (4th ed.). OECD Publishing. <https://doi.org/10.1787/9789264304604-en>
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211–218. <https://doi.org/10.1145/505248.506010>
- Prochnow, D. A., Vieira, E. P., Baggio, D. K., & Thesing, N. J. (2025). Governança colaborativa em um ecossistema regional de inovação: O caso do programa Inova RS. *Revista Gestão & Planejamento*, 26, 310–326. <https://revistas.unifacs.br/index.php/rgb/article/view/9397>

- Retterath, A., & Braun, R. (2020). Benchmarking venture capital databases [Working paper]. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3706108
- SICT-RS. (2026). Rede RS Startup. Secretaria de Inovação, Ciência e Tecnologia do Rio Grande do Sul. <https://sict.rs.gov.br/rede-rs-startup-6a4fb734494f5>
- Stam, E. (2015). Entrepreneurial ecosystems and regional policy: A sympathetic critique. *European Planning Studies*, 23(9), 1759–1769. <https://doi.org/10.1080/09654313.2015.1061484>
- Stam, E., & van de Ven, A. (2021). Entrepreneurial ecosystem elements. *Small Business Economics*, 56(2), 809–832. <https://doi.org/10.1007/s11187-019-00270-6>
- van Aken, J. E. (2004). Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological rules. *Journal of Management Studies*, 41(2), 219–246. <https://doi.org/10.1111/j.1467-6486.2004.00430.x>
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33. <https://doi.org/10.1080/07421222.1996.11518099>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.1080/01621459.1927.10502953>
- Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* (pp. 354–359). American Statistical Association.
- Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1), 41–63. <https://doi.org/10.1111/j.1467-9574.2011.00508.x>

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença [Creative Commons CC-BY](#).
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.