How Much Local Data Is Enough? An Ontology-Grounded Neurosymbolic Research Agenda for Transferability in Forecasting

Rodrigo Huber Marques Moreira Mendes[1,*], Fernanda Baião[1] and Reinaldo Castro Souza[1]

[1]*Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Marquês de São Vicente St. 225, Gávea, Rio de Janeiro, 22451-900, Brazil*

Abstract

Renewable-energy planning depends on reliable solar-irradiation forecasts, yet measurement infrastructure remains sparse across much of Brazil, leaving practitioners with abundant data at a few well-instrumented sites and almost nothing elsewhere. This paper addresses that scarcity through a *neurosymbolic* (NSAI) approach: rather than treating transfer learning as a purely statistical operation, I propose to embed a well-founded domain ontology inside the decision procedure that should govern whether a model trained at a data-rich source site may be deployed at a data-scarce target site, and if so, how much local data the target needs before adaptation is warranted. The central question is partly conceptual and partly empirical: *when can a model trained in one context be trusted in another, and how much local evidence is needed before that trust is justified?* Within this paradigm, symbolic, connectionist, and statistical components are combined; the distinguishing choice here is to ground the symbolic component in a well-founded domain ontology, which would formalize domain knowledge about site similarity (climate type, biome, geography, radiative regime) so that ontological commitments become explicit, inspectable, and testable. To this end, I am developing a *Neurosymbolic Semantic Gate*: a symbolic decision layer that will operationalize these commitments to govern transfer. This statement lays out the motivation, the core ideas, and the experimental program: a principled investigation of when ontological structure can govern transfer, and of how much local data a target site truly needs.

Keywords

Transfer Learning, Neurosymbolic AI, Ontology, data efficiency, Solar Irradiation, Forecasting, Machine Learning, Renewable Energy

CEUR-WS.org/Vol-4342/short2.pdf

1. Motivation: A Practical Question with Ontological Teeth

Renewable-energy planning needs reliable forecasts of solar irradiation, and good forecasts need data. In Brazil, a continent-sized country with a sparse and uneven measurement network, that requirement is rarely met. A handful of well-instrumented sites have years of clean records; vast regions have almost nothing. The pragmatic hope is *transfer learning* [1]: train a model where data is plentiful and reuse it where it is not. Yet reuse is not epistemically neutral: it rests on the assumption that the target location is sufficiently similar to the source context in the aspects that govern predictive performance. It presupposes, in other words, that the target is in some relevant respect *the same kind of place* as the source. This similarity requirement is well recognized in the transfer-learning literature [1, 2]: when source and target domains diverge along dimensions that matter, transfer can degrade performance rather than improve it (*negative transfer*).

That presumption is exactly an ontological claim. "The same kind of place" is not a statistical primitive; it is a judgment about which properties of a site are essential to its solar behavior and which are incidental. Two stations may share a latitude but differ in cloud regime; share a biome but differ in altitude; sit far apart yet behave identically. Deciding which similarities warrant transfer, and which are mirages, is a question formal ontology was built to clarify [3]. An informal or inherited classification

Proceedings of FOIS 2026 Satellite events co-located with the 16th International Conference on Formal Ontology in Information Systems (FOIS 2026), September 21-25, 2026, Vitória, Brazil.

*Corresponding author.

rodrigo@aluno.puc-rio.br (R. H. M. M. Mendes); fbaiao@puc-rio.br (F. Baião); reinaldo@puc-rio.br (R. C. Souza)

0009-0003-3001-2870 (R. H. M. M. Mendes); 0000-0001-7932-7134 (F. Baião); 0000-0002-0210-3703 (R. C. Souza)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-27

(such as a Köppen map or a biome scheme) can supply candidate categories, but it cannot, on its own, make those commitments explicit, distinguish essential from accidental properties, or specify the conditions under which a similarity judgment is defeasible. A well-founded domain ontology is needed precisely to discharge that role: to articulate, in a logically precise and scientifically grounded vocabulary, what it means for two sites to be *the same kind of place* for the purpose of inductive reuse, and to make that articulation testable. My doctoral project takes this seriously: it treats *transferability* not as an engineering heuristic whose requirements can be met by purely empirical methodology, but as an ontological problem whose structure can be investigated by making domain commitments explicit and open to empirical test.

The research question is simple to state and hard to answer: **How much local data is enough?** Given a source model and a target site, how much local data (spanning whichever dimensions of variation are ontologically relevant, including temporal horizon, altitude range, cloud-regime diversity, and seasonal coverage) must be collected before fine-tuning the source model on target observations reliably outperforms deploying it as-is? Call that quantity the site's *data budget*, n^*. A theory of n^* would help practitioners estimate, before investing in instrumentation, how much data is likely to be required and would also clarify what governs the portability of learned knowledge.

2. Intellectual Roots: Neurosymbolic AI and the Role of Well-Founded Ontology

The conceptual spine of my agenda is the renaissance of *neurosymbolic AI* (NSAI): the recognition that symbolic, connectionist, and statistical approaches are complementary rather than competing, and that coupling them enables governance properties that neither achieves alone [4, 5]. Within that broad paradigm, the choice of *what* populates the symbolic layer is decisive. My contribution is to ground that layer in a *well-founded domain ontology* [6], so that the symbolic knowledge it uses is structured, semantically precise, inspectable, and testable.

The Governance Promise of NSAI. Purely statistical models are powerful but mute about their own scope: a neural network does not know *where* it applies. The NSAI response is to couple the learning component with an explicit symbolic layer that carries structured, inspectable knowledge and can be interrogated in human-legible terms [4, 5]. The promise is not only better accuracy but also *governance*: a model whose limits can be stated in plain terms, including where it should and should not be trusted. In my setting, the learner is a source forecasting model, and the symbolic layer is a gate, still under development, that will evaluate whether a target site is an appropriate deployment context and how much local evidence is needed before transfer from source to target is justified.

Well-Founded Ontology as the Differentiator. I propose to ground the symbolic layer in a *well-founded domain ontology* rather than an ad-hoc feature set. A symbolic gate is only as good as the categories it reasons with, and the categories here ("same climate," "same biome," "ecologically near") are inherited from scientific classification systems (Köppen climate types [7, 8], biome schemes [9], bioclimatic distances [10]) that are, in effect, domain ontologies: structured vocabularies for organizing the world in scientifically meaningful ways. Grounding the gate in such ontologies makes its symbolic commitments semantically precise and independently auditable [11]. It also raises a genuinely ontological question that the FOIS community is well-placed to sharpen: *which* notion of sameness is the *right* one for a given inductive purpose? Climate identity, ecological proximity, and geographic nearness are distinct relations that often disagree. A theory of transferability must specify which of these relations, if any, has predictive relevance for source-to-target knowledge transfer.

My research sits at this intersection within the NSAI paradigm. The neurosymbolic architecture gives the ontological commitments a concrete job to do; the empirical study then *tests* those commitments against measured data budgets. Ontology generates hypotheses about source–target similarity, and transfer experiments adjudicate them. The project is therefore more than a standard machine-learning

application: it offers a setting for studying how ontologically grounded similarity relates to inductive performance, and whether well-founded categories carry more predictive value than operational, continuous proxies.

3. Core Idea: A Neurosymbolic Semantic Gate

Context. Transfer learning offers a practical path around data scarcity: train a forecasting model where observations are plentiful and reuse it where they are not. Yet a neural network is mute about the scope of its own validity: it carries no account of *why* or *where* learned patterns should generalize. Bridging that gap requires coupling the statistical learner with an explicit, inspectable symbolic layer that reasons in domain terms about source–target compatibility.

Research Problem. The key question is which notions of source–target similarity actually predict whether a source model will transfer to a target site, how quickly it will transfer, and how much local data the target site needs before fine-tuning reliably beats the zero-shot baseline. Existing transfer-learning practice treats similarity heuristically; no principled account explains why certain criteria (climate class, biome proximity, geographic nearness) should have inductive weight, nor adjudicates among them empirically.

Hypotheses. The study is organized around two claims I intend to put under adversarial test, not to assume:

1. **(Ordering.)** If ontological similarity (as operationalized by the gate) orders source–target pairs, then sites judged more similar to the source require fewer local observations to adapt, yielding a measurably lower data budget n^*.
2. **(Discrimination.)** If the candidate notions of sameness are *not* interchangeable, then some (such as categorical climate identity) will carry systematically different inductive values from others (such as graded behavioral proximity), and the empirical record will differentiate them.

Stating the hypotheses precisely enough to be falsifiable is itself part of the contribution: similarity heuristics are often said to "work" without being tested under adversarial conditions, and I want these criteria to be judged under that kind of scrutiny.

Proposed Solution. The proposed solution is a *Neurosymbolic Semantic Gate*: a symbolic decision layer that sits between a trained source model and any prospective target site, and that reasons, transparently in terms a domain scientist can read, about whether and how readily transfer is warranted. The gate is designed as a cascade of symbolic criteria, each encoding a distinct ontological notion of source–target similarity and contributing a human-readable judgment to an overall verdict (APPROVED, WARNING, or REJECTED), perhaps with a supporting confidence level, together with an uncertainty estimate over the predicted data budget. Each criterion is a *distinct hypothesis* about what makes two sites the same kind of place for the purpose of solar forecasting (Table 1). The design keeps these notions separate so the study can test a more specific question: which notion of sameness, if any, actually carries predictive weight?

Evaluation Methodology. The gate will be evaluated against an empirically measured data budget n^*, the smallest amount of local data after which a fine-tuned model consistently outperforms the zero-shot source model. This reference budget has already been established across 20 Brazilian sites: one source site (Galeão, Rio de Janeiro) and 19 target sites spanning roughly 11 to 2,100 km, from the tropical Rio de Janeiro basin to the semi-arid north, using reanalysis weather data [12]. Each criterion in Table 1, together with the gate's aggregate verdict, will then be assessed for its ability to rank target sites by n^* and to identify which notion of sameness does the predictive work. The analysis is designed to control for target-site difficulty, so that any apparent effect of similarity is not an artifact of a site being inherently easier to forecast.

Table 1
Candidate notions of source–target similarity to be encoded by the gate. Each is a distinct, testable hypothesis about what licenses transfer; the project is designed to discriminate among them rather than assume any one in advance.

Criterion	Notion of "sameness"	Ontological source
Radiative compatibility	Comparable top-of-atmosphere forcing	Astronomy / site geometry
Seasonal alignment	Matching shape of the annual cycle	Climatology
Climate identity	Same Köppen class	Köppen–Geiger classification [7, 8]
Ecological proximity	Near in biome / climate space	Whittaker biome scheme [9]
Geographic nearness	Small great-circle distance	Spatial proximity

4. Research Plan

The agenda runs from mid-2026 to the projected thesis defense in July 2027 (roughly months M1–M13 below), organized into three overlapping work packages rather than a strictly sequential set of phases.

(WP1, M1–M3) Measure the Budget. For all 20 sites, the data budget n^* defined in Section 3 has been measured under a fixed protocol. This work package, already largely complete, delivers the ground-truth reference against which the gate's symbolic reasoning will be evaluated.

(WP2, M2–M9) Formalize and Build the Ontology-Grounded Gate. The well-founded domain ontology and the symbolic gate that operationalizes it are developed here: clarifying *what* the system commits to when it calls two sites "similar" (separating the roles a site plays, since source and target are relational roles rather than intrinsic types; the graded relations that mediate them; and the epistemic status of a verdict, a defeasible judgment rather than a fact about the world), and encoding those commitments as the cascade of criteria in Table 1. This is the most demanding package and the one where I most expect to draw on, and contribute to, the formal-ontology literature [6].

(WP3, M7–M12) Analyze the Gate against Measured Budgets. This package carries out the evaluation defined in Section 3, assessing each criterion and the gate's aggregate verdict against the measured budgets to test the Ordering and Discrimination hypotheses. Its results directly validate or falsify the central claims of the project ahead of the final write-up.

5. Research Vision

The thesis is that **transferability is a structured, investigable phenomenon, and that formal ontology has a concrete, testable role in explaining it**, not as decorative support for a machine-learning pipeline, but as a source of hypotheses that a learning system can evaluate. That thesis rests on methodological presuppositions that are themselves open to scrutiny.

First, the inherited scientific classifications the gate will draw on (Köppen climate types [7, 8], Whittaker biome schemes [9, 13], bioclimatic distance measures [10]) encode ontological commitments about what makes two environments alike; those commitments may or may not carry inductive value for a given forecasting task, and the project treats this as an empirical question rather than a design assumption. Second, a gate verdict is intended as a *defeasible epistemic judgment*: the best conclusion licensed by available categorical evidence, not a fact about whether a target site will respond well to transfer; new local observations can and should defeat it. Third, for ontological commitments to be scientifically useful they must be operationalized precisely enough to yield predictions that are, in principle, falsifiable; vague appeals to "similarity" are insufficient. Fourth, the evaluation adopts an adversarial stance: the experimental protocol is designed not merely to demonstrate that the gate's verdicts are sometimes correct, but to identify the conditions under which it may fail. A standard of

acceptance that has never been challenged is indistinguishable from an assumption; adversarial testing is what transforms a working criterion into a validated one.

Three commitments organize the agenda:

1. **Ontology as testable hypothesis, not assumed scaffolding.** Every claim that "X-similarity licenses transfer" should be stated precisely enough to be testable, and then evaluated empirically. I want a research practice in which ontological commitments are not quietly assumed but openly adjudicated by evidence.
2. **The epistemology of similarity.** A question I want to pursue beyond solar forecasting is when inherited scientific classification carries inductive value, and when similarity must instead be re-grounded in operational, continuous terms. This matters to formal ontology itself, because it concerns the relationship between our category systems and the inferences they are meant to support.
3. **Neurosymbolic governance with an accountable symbolic layer.** The long-term vision is learning systems whose symbolic component provides inspectable governance, indicating where a model may be trusted and how much local evidence a new deployment needs, while remaining accountable to data. A symbolic layer that makes no checkable commitments cannot be trusted; one whose commitments are tested and survive can. The vision is to build the second one.

Beyond the solar case, this same skeleton (a source model, a new context, a local adaptation budget, and competing similarity relations that may predict that budget) appears in other scientific forecasting settings, including hydrology, air quality, agriculture, and almost all types of renewable energy sources, as well as in any domain where models are redeployed under data scarcity. A more general methodological contribution is a transferable *way of asking*: state the ontological commitment, assign it a role within a neurosymbolic system, and let the empirical outcome adjudicate.

6. Why FOIS, and What I Seek from the ECS

The research question at the center of this proposal (**which ontological notion of similarity governs knowledge transfer, and how much local evidence is required before transfer is justified**) is one node in a broader doctoral research plan whose overarching question is how formal ontology can make neurosymbolic systems both principled and empirically accountable. The solar-forecasting transfer-learning setting was chosen because it is my field of expertise; that familiarity, together with the formal-ontology and neurosymbolic framing, makes the broader question concrete and tractable: a gate's verdicts can be measured against observed data budgets, the similarity relations are drawn from established domain ontologies, and the experimental design is reproducible. It is precisely because this node of the plan sits at the intersection of inductive learning and formal ontology that I am bringing it to FOIS rather than to a machine-learning venue: the crux is not the forecasting model but the question of which notions of *sameness* have the standing to govern inductive reuse, a question I cannot answer well without the conceptual tools and the critical community that FOIS represents. From the Early Career Symposium, I am seeking critical and constructive feedback that sharpens a project: pushback on how I individuate and operationalize similarity relations, on the epistemic status I assign to a gate's verdict, and on how to frame the interface between ontology and learning so that it speaks to both communities. The mentoring and fast-talk formats provide an appropriate setting to test the proposal and connect with researchers whose work informs it.

Acknowledgments

This research is conducted at PUC-Rio, Department of Industrial Engineering, under the supervision of Professor Fernanda Baião and Professor Reinaldo Castro Souza, whose guidance shaped the questions

addressed in this statement. I thank the FOIS Early Career Symposium organizers and mentors for the opportunity to present this work. Also, gratefully acknowledge the CAPES, CNPq, and FAPERJ research agencies for their financial support.

Declaration on Generative AI

During the preparation of this work, the authors used a language model assistant for LaTeX mathematical writing, language refinement, and orthography checking. The authors reviewed and edited the manuscript and take full responsibility for its content.

References

[1] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (2010) 1345–1359. doi:10.1109/TKDE.2009.191.

[2] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big Data 3 (2016) 9. doi:10.1186/s40537-016-0043-6.

[3] N. Guarino, D. Oberle, S. Staab, What is an ontology?, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, 2nd ed., Springer, Berlin, Heidelberg, 2009, pp. 1–17. doi:10.1007/978-3-540-92673-3_0.

[4] A. d. Garcez, L. C. Lamb, Neurosymbolic AI: the 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406. doi:10.1007/s10462-023-10448-w.

[5] H. Kautz, The third AI summer: AAAI robert s. engelmore memorial lecture, AI Magazine 43 (2022) 105–125. doi:10.1002/aaai.12036.

[6] G. Guizzardi, Ontological Foundations for Structural Conceptual Models, Ph.D. thesis, University of Twente, Telematica Instituut / CTIT, 2005.

[7] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, E. F. Wood, Present and future Köppen–Geiger climate classification maps at 1-km resolution, Scientific Data 5 (2018) 180214. doi:10.1038/sdata.2018.214.

[8] M. C. Peel, B. L. Finlayson, T. A. McMahon, Updated world map of the Köppen–Geiger climate classification, Hydrology and Earth System Sciences 11 (2007) 1633–1644. doi:10.5194/hess-11-1633-2007.

[9] R. H. Whittaker, Communities and Ecosystems, 2nd ed., Macmillan, New York, 1975.

[10] S. E. Fick, R. J. Hijmans, WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology 37 (2017) 4302–4315. doi:10.1002/joc.5086.

[11] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, Artificial Intelligence 302 (2022) 103627. doi:10.1016/j.artint.2021.103627.

[12] R. Gelaro, et al., The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), Journal of Climate 30 (2017) 5419–5454. doi:10.1175/JCLI-D-16-0758.1.

[13] T. Conradi, J. A. Slingsby, G. F. Midgley, H. Nottebrock, A. H. Schweiger, S. I. Higgins, An operational definition of the biome for global change research, New Phytologist 227 (2020) 1294–1306. doi:10.1111/nph.16580.