Open Knowledge for a Sustainable Future: Research, Ethics, and Wikipedia

Week 5 — FAIR Principals - Multilingual and Ethical considerations

Francis Bond (**Academic**) Pavel Bednařík (Wiki)

Palacký University Olomouc | Wikimedia ČR

14 October 2025

Roadmap

Why FAIR Data?

- **Findable** — So that other researchers (and future you) can actually discover your data.
  If no one can find it, it might as well not exist.
- **Accessible** — So that once found, data can be retrieved easily and safely.
  Access should be possible even years later, with clear rules if restrictions apply.
- **Interoperable** — So that data from different projects can be combined or compared.
  Shared formats and vocabularies let computers and people understand each other.
- **Reusable** — So that data can be meaningfully used beyond its original purpose.
  Good documentation and clear licensing allow others to build on your work.

FAIR data makes research more transparent, verifiable, and sustainable.
Sources: (Wilkinson et al., 2016), FAIR Data 101 (accessed 2025-11-08)

Findable: why it matters

- Data has little value if no one can locate it.
- Clear titles, metadata, and persistant identifiers (like DOIs) make your work visible to both people and search engines.
- When data is easy to find, it is easier to cite — increasing your impact and ensuring credit.

Findable is putting your data on the map.

Making data findable

- Use descriptive, consistent names for datasets and files.
- Add rich metadata — who, what, when, where, why.
  - Use standard terms, so people can find them easily
- Register your dataset in a trusted repository with a permanent identifier (e.g., Zenodo, Dataverse).
- Cross-link: paper ↔ dataset ↔ code.
- This applies to governments and companies as well as academia
  - Making data findable should be part of making the data!

Accessible: why it matters

- Even well-described data is useless if people can't reach it.
- Accessibility ensures that others can actually download or view your work — now and in the future.
- It also clarifies boundaries: who can use the data, under what conditions.
-

Access is about clarity and stability, not about giving everything away. For sensitive language/cultural data, clarity of access rules is part of the ethics record.

Making data accessible

- Store your data in repositories that use open web standards (HTTP/HTTPS).
  - Use persistant repositories
  - Support persistant repositories
  - Data from companys is typcially very perishable
- Keep metadata available even if files must be restricted.
- Explain how others can request access if necessary.
- Test from a new computer — can someone else follow your instructions?

Interoperable: why it matters

- Science advances when we can combine data from many sources.
- Interoperability allows your work to connect with others'
  — across languages, disciplines, and tools.
- Shared standards prevent "data silos" that only one group can use.

Interoperable data talks easily to other data.

Making data interoperable

- Use open, non-proprietary formats (CSV, JSON, XML).
  - Text is better than binary, as you can always read it
- Follow community conventions for terminology and encoding.
- Provide clear documentation of structure and meaning.
- Link to external vocabularies or identifiers when possible.
  - e.g. ISO 639 for the representation of languages and language groups.
    - English: en, eng (ISO 639-1, 2&3)
    - Czech: cs, cze/ces, ces (ISO 639-1, 2-B/T, 3)

Reusable: why it matters

- Good data should outlive the project that created it.
- Reuse lets others verify results, test new questions, and save time.
- Well-documented and licensed data reduce duplication and waste.
- Producing data is the job of many agencies, but if no one uses it, it is wasted effort, …

Reuse is the payoff — the reward for all your FAIR effort.

Making data reusable

- Include full context: how the data was collected, cleaned, and analysed.
- Choose a clear license (e.g., CC BY or CC BY-SA).
- Describe known limitations or ethical constraints.
- Provide examples of how to cite or acknowledge your dataset.

Citing the Open Multilingual Wordnet

If you use these wordnets, please cite the original projects who created them, if you got value from this aggregation/normalization, please cite Bond and Paik (2012).

References

Francis Bond and Kyonghee Paik (2012) A survey of wordnets and their licenses In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71

Open licences and the Open Definition

- **Open data** means more than free access — it means legal permission to *use, modify, and share* without discrimination.
- The **Open Definition** (Open Knowledge Foundation) states that data is open if "anyone can freely access, use, modify, and share for any purpose" — subject only to requirements of attribution and share-alike.
- Common open licences:
  - **CC BY** – reuse with attribution.
  - **CC BY-SA** – reuse with attribution and same licence (share-alike).
  - **CC0 /Public Domain** – no restrictions.
  - **ODC BY/ODbL** – for databases; require attribution and share-alike.
- Choosing the right licence ensures that your data remains reusable and legally safe.

Openness is a design choice—licensing makes it possible, clarity makes it trustworthy.

When not to use open licences

- Openness is a virtue, but not a universal rule.
- Some data should *not* be released under open licences because openness could cause harm.
- Examples include:
  - Personal or medical information that identifies individuals.
  - Cultural or linguistic materials shared under community protocols.
  - Locations of endangered species or sacred sites.
  - Data collected without full, informed consent for public reuse.
- In such cases, use restricted or tiered access, or licences that reflect community agreements.

Responsible openness means balancing transparency with care, consent, and context.

Roadmap

CARE: what and why

- **FAIR** focuses on making data easy to find and reuse.
- **CARE** ensures that openness respects people, communities, and context.
- Developed by the **Global Indigenous Data Alliance (GIDA)** (Carroll, Garba, et al., 2020).
  - **Collective Benefit:** Who gains from sharing this data?
  - **Authority to Control:** Who decides how it can be used?
  - **Responsibility:** Are researchers accountable to those represented?
  - **Ethics:** How can openness coexist with respect and consent?

CARE asks: who benefits, who decides, and how can openness support justice?

Together, FAIR + CARE promote data that is both open and just.

CARE – Collective Benefit

- **Why it matters:** Data can empower—or exploit—communities.
- Sharing should create positive outcomes for the people represented, not just for researchers or institutions.
- Benefits may include
  - knowledge preservation
  - recognition
  - capacity building
  - economic return

Example

A linguistic corpus created with a minority language group should support local education, not only external research.

Open data should contribute to shared well-being, not just extraction.

CARE – Authority to Control

- **Why it matters:** Communities have the right to govern how their cultural data is used and shared.
- Data sovereignty recognises that Indigenous and local groups decide access levels, licensing, and reuse.
- Authority can be expressed through traditional protocols, community review, or legal agreements.

Example

Before releasing ethnobotanical recordings, obtain explicit approval from the community council and document their terms of use.

Openness must be chosen, not imposed.

CARE – Responsibility

- **Why it matters:** Data stewardship is a relationship, not a transaction.
- Researchers have continuing responsibilities to maintain context, accuracy, and cultural respect.
- Responsibility includes acknowledging contributors, correcting misuse, and ensuring long-term protection.

Example

Maintain communication with data contributors; report how their data are reused and share new findings with them.

Responsible openness strengthens trust and collaboration.

CARE – Ethics

- **Why it matters:** Ethical reflection must guide decisions about collection, sharing, and reuse.
- Even technically FAIR data can be unethical if obtained without consent or if it exposes communities to harm.
- Ethics involves empathy, transparency, and accountability throughout the data lifecycle.

Example

If publishing traditional songs or stories, consider whether public release might violate cultural norms or sacred boundaries.

Ethics is the foundation that keeps openness humane.

CARE beyond Indigenous contexts

- The CARE principles were developed for Indigenous and community-governed data.
- Yet their spirit—centering people, consent, and benefit—applies to many creative and online communities.
- Think of fanfiction, collaborative writing, or online art:
  - **Collective Benefit:** Sharing enriches a fandom; readers and writers learn from one another.
  - **Authority to Control:** Authors decide if and where to post, and may lock or delete works.
  - **Responsibility:** Platforms and moderators ensure respect and prevent harassment.
  - **Ethics:** Respect creators' boundaries, content warnings, and attribution norms.
- These are not the original CARE cases—but they show how ethical principles travel.

CARE is a mindset: who benefits, who decides, who is responsible, and who might be harmed?

FAIR, CARE, and sustainability

- Openness is not only ethical — it is environmental.
- Reusing well-documented data prevents repeated fieldwork, duplicated storage, and wasted energy.
- **FAIR** encourages efficient data management: organised, shareable, easier to maintain.
- **CARE** reminds us that sustainable research also means social sustainability — maintaining trust, reciprocity, and respect.
- Together, FAIR + CARE reduce both the carbon and the ethical footprint of digital scholarship.

Sustainability begins when every dataset collected is used to its fullest and shared with care.

Towards green open research

- Each stage of the data lifecycle has environmental costs: storage, transfer, and computation.
- Good FAIR practice—clean metadata, smaller files, open formats—cuts energy use and prolongs data life.
- CARE principles promote community partnerships that reduce extractive, short-term projects.
- Reproducible workflows and open infrastructure avoid unnecessary duplication of effort and computation.

Practical takeaway

When you make data open, make it *efficiently* open: small, meaningful, reused, and responsibly hosted.

Sustainable openness means doing more with less—knowledge without waste.

Roadmap

AO3: an open, community archive

- **Archive of Our Own (AO3)** is a non-commercial platform for fanfiction and other transformative works.
- Built and maintained by the **Organization for Transformative Works (OTW)**, a volunteer-run nonprofit.
- Serves millions of readers and writers; demonstrates how openness can align with community values.

AO3 shows that open infrastructure can succeed ethically when the community defines the rules.

AO3 and the FAIR Principles

- AO3's design reflects key FAIR ideas: findability, accessibility, interoperability, and reusability.
- Its success depends on both *good metadata* and *good governance*.
- Each user contribution becomes part of a well-organised, searchable cultural dataset.

Findable and Accessible

- **Findable:** AO3 uses detailed, user-generated tags and filters. Works can be discovered by fandom, pairing, trope, or content rating.
- Every work has a stable URL and consistent metadata, ensuring it can be cited or referenced reliably.
- **Accessible:** Content is available through a standard web interface and open API.
- Authors may choose to “lock” works for logged-in users—clear, transparent access rules.

Users can always find and retrieve what they are looking for, within consent boundaries.

Interoperable and Reusable

- **Interoperable:** AO3 maintains a controlled vocabulary of tags with hierarchical relations—useful for digital humanities and fan studies.
- Data can be exported via API, allowing analysis across fandoms or comparison with other archives.
- **Reusable:** The platform supports transformative reuse—the foundation of fan creativity.
- Clear terms of service define attribution and remix norms, promoting responsible reuse.

AO3 enables a living ecosystem of reuse—stories inspire stories.

AO3 and the CARE Principles

- AO3 also embodies the CARE principles: Collective Benefit, Authority to Control, Responsibility, and Ethics.
- Its governance structure protects authors' rights and ensures collective well-being.
- CARE explains *how* openness can exist without exploitation.
  - **Collective Benefit:** AO3 preserves fan culture, ensuring benefits return to its community rather than corporations.
  - **Authority to Control:** Authors retain ownership and can delete or restrict access at any time.
  - **Responsibility:** OTW provides clear policies, content warnings, and volunteer training.
  - **Ethics:** Inclusivity and consent are central; moderation prevents harm while protecting creative freedom.

Lessons from AO3

- FAIR and CARE can coexist: openness does not mean loss of control.
- Community-defined metadata and tagging ensure technical quality *and* social relevance.
- Respectful governance and transparency sustain participation and trust.
- AO3 demonstrates that ethical, participatory data stewardship is both possible and scalable.

AI training and the CARE principles

- Modern AI systems are trained on massive text, image, and audio datasets—often collected without explicit consent.
- This raises serious ethical and governance questions:
  - Who owns or controls the data used to train AI?
  - Who benefits from the commercial and social value it generates?
  - How are communities represented—or misrepresented—in AI outputs?
- The **CARE Principles** provide a useful lens:
  - **Collective Benefit:** Who gains from AI trained on public or cultural data?
  - **Authority to Control:** Were contributors able to consent or opt out?
  - **Responsibility:** Are data origins and attributions transparent?
  - **Ethics:** Does model training cause harm or reinforce bias?

Responsible AI should provide not only FAIR access to data, but CARE for the people behind it.

Examples and ongoing research

- Scholars are exploring how CARE can inform AI ethics and governance:
  - **Carroll et al. (2020)** – Introduced CARE as complement to FAIR, warning that AI amplifies inequities.
  - **Whittaker et al. (2023)** – Frame AI data practices as extractive and colonial.
  - **Bender et al. (2021)** – "Stochastic parrots" critique: AI models reuse data without consent or understanding.
  - **DECORAIT (2023)** – A registry to track opt-in/opt-out for creative data in AI training.
- CARE calls for:
  - Transparent dataset documentation (e.g., "model cards," "data statements").
  - Mechanisms for consent, credit, and benefit-sharing.
  - Recognition that "open" $\neq$ "free to exploit."

Sources: (Bender et al., 2021; Carroll, Hudson, et al., 2020; Smith et al., 2023; Whittaker et al., 2023).

Generative AI: opportunities and risks

Potential benefits

- **Access:** Makes information, examples, and creative ideas instantly available.
- **Efficiency:** Supports brainstorming, translation, and summarisation.
- **Inclusion:** Lowers barriers for non-native speakers or early-career writers.
- **Education:** Helps students explore styles, genres, and feedback interactively.

Key concerns

- **Accuracy:** Output may be plausible but factually wrong or biased.
- **Attribution:** Original authors of training data are rarely credited.
- **Motivation:** If creators lose recognition, incentive to write or share decreases.
- **Transparency:** Training data sources and usage policies often undisclosed.
- **Sustainability:** Large models consume significant energy and resources.

Generative AI amplifies the FAIR goals of access and reuse—but challenges the CARE goals of credit, consent, and responsibility.

Reading, attribution, and true knowledge

- Responsible scholarship begins with **reading the sources themselves**.
- To cite ethically, you must:
  - **Read** the original work in full.
  - **Understand** its arguments and evidence.
  - **Judge** its credibility.
  - **Attribute** it accurately, giving credit to the authors and allowing others to check your conclusions.
- AI-generated summaries or paraphrases skip these steps:
  - You cannot check what was included, omitted, or distorted.
  - You lose the ability to evaluate the quality and reliability of the evidence.
  - As a result, genuine knowledge building—where ideas build on verified understanding—becomes impossible.
- **Reading is thinking.** Without it, citation becomes mimicry, not scholarship.
- **Writing is thinking.** Without it, you cannot build your own connections.

Roadmap

Persuasion and ethics in writing

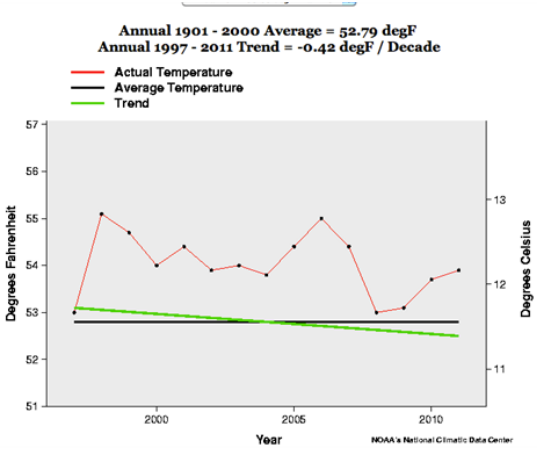
- All academic writing is persuasive: we choose what to show, emphasise, or omit.
- **Ethical persuasion** means presenting evidence fairly and reasoning transparently.
- Our aim is to convince by clarity and integrity—not by manipulation.
- The same ethics that guide FAIR and CARE also guide good writing: accuracy, context, respect.

If the data is sound but our argument misleads, we still fail ethically.

Common pitfalls in persuasive research writing

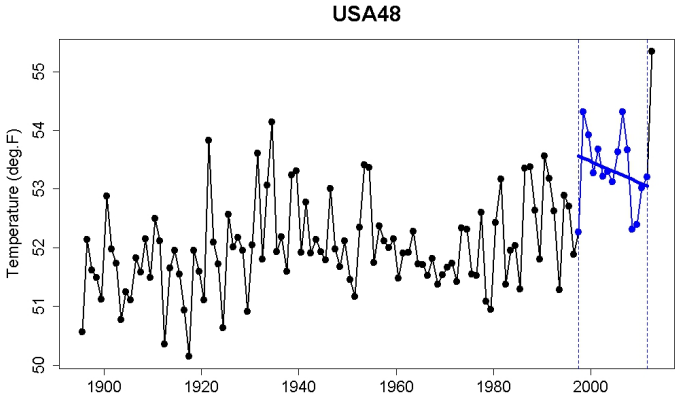
- **Cherry-picking data:** presenting only results that fit your argument.
- **Overstating causality:** “X causes Y” when evidence shows only correlation.
- **Ignoring confounders:** neglecting variables that explain the pattern.
- **Ambiguous visuals:** missing labels, unclear units, or distorted scales.
- **Rhetorical exaggeration:** emotional or absolute language (*proves, always, never*).

Cherry-picking data: The temperature is declining!



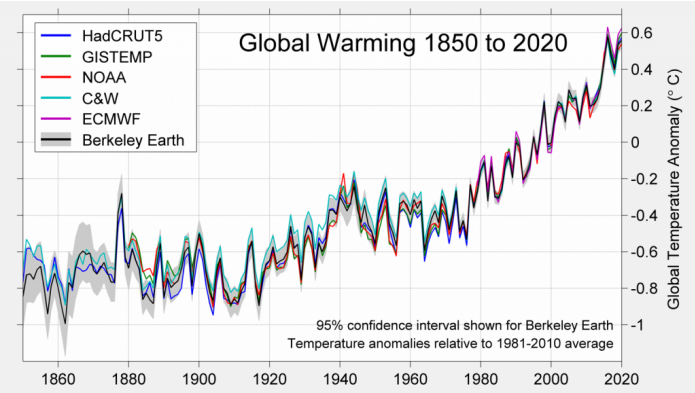
John Coleman, cited by Anthony Watts, cited by Open Mind

Not if you zoom out



Open Mind *Cherry Picking is Child's Play*

And average over different regions



Climate Change Commission *CCC: No more time for inaction as global warming accelerates, marking 2020 as one of the warmest years on record*

Example: cherry-picking data

What was done wrong?

- Only the years supporting the argument were shown.
- Broader data contradicts the claimed trend.
- No context or uncertainty provided.

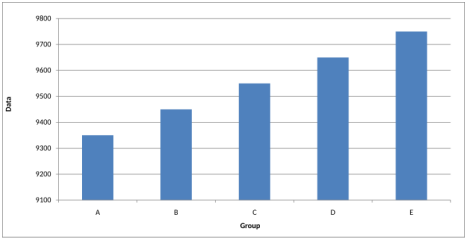
Better practice:

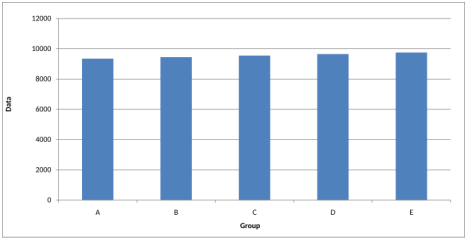
- Show full dataset and acknowledge anomalies.
- Explain possible causes rather than hiding them.
- Show confidence intervals (the grey band in the last image)

If someone cherry picks data, and refuses to acknowledge it after it has been pointed out, then they are not a credibale source, …

How (not) to lie with graphs

- **Truncated axes** exaggerate small changes.
- **Omitted error bars** hide uncertainty.
- **3D charts** distort visual proportions.
- **Unequal scales** misrepresent relationships.





Images from Wikipedia:Misleading graph (accessed 2025-11-08)

From data to argument: keeping FAIR and CARE alive

- **FAIR failures**—missing metadata, unclear methods—lead to weak or unverifiable claims.
- **CARE failures**—ignoring context or consent—lead to distorted or harmful narratives.
- Ethical writing means:
  - Verify data quality and provenance before citing.
  - Present uncertainty honestly.
  - Represent people and communities with respect.
- If we manage data carelessly, our communication cannot be trusted.
- This is as true for citing writing as it is for citing data!

Persuasion with integrity

- Ethical persuasion is clear, evidence-based, and proportionate.
- Data visualisation should illuminate, not impress.
- Sound governance (FAIR/CARE) leads to sound communication.
- Our goal should be **to inform, not to manipulate.**

Roadmap

When what you find surprises you

- You hypothesised X — but your data support Y, or nothing at all.
- This is not failure — it is how discovery happens.
- Many great findings began as “wrong” results that researchers chose to explore rather than hide.
- Ethical research means reporting what you found, not what you hoped to find.

If you discover that what you thought was wrong, *then you have improved your model of the world* — if you just confirm what you thought, you have gained less!

Ethical responses to unexpected results

- **Reflect:** Why might this have happened? Re-examine assumptions, methods, or definitions.
- **Consider alternatives:** Could sampling, measurement, or confounding factors explain the result?
- **Report transparently:** Document what changed—methods, data, or interpretation.
- **Revise hypotheses:** Update theory in light of evidence rather than forcing the evidence to fit theory.
- **Stay curious:** Unexpected outcomes often open richer questions.

FAIR + CARE: supporting honest revision

- **FAIR** practices—clear metadata, versioning, provenance—let others trace how results emerged.
- **CARE** principles—context, consent, responsibility—help ensure surprises are interpreted respectfully.
- Keeping complete records allows you (and others) to:
  - Reanalyse the same dataset later.
  - Identify methodological errors.
  - Learn from data collected by others with proper context.
- This applies to *found data* too: when reusing open datasets, respect their limits and document your reasoning.

Roadmap

Scientific, creative, and encyclopedic writing

- **Scientific writing** — formal, objective, structured
  - Evidence-based, peer-reviewed, reproducible.
  - Goal: to inform and test ideas transparently.
- **Creative writing** — expressive, narrative, open to emotion and metaphor.
  - Goal: to move, provoke, or entertain.
  - Truth is emotional or imaginative, not empirical.
- **Wikipedia writing** — concise, neutral, verifiable.
  - Summarises consensus rather than taking sides.
  - Accessible to a broad, multilingual audience.

Ethical imagination

- Fiction gives extraordinary freedom — should there also be responsibility?
- Writers shape how readers see the world, other cultures, and themselves.
- **Ethical imagination** means using creativity without exploiting others' pain or identity.
- Ask:
  - Whose stories am I telling?
  - Do I have the right, or the context, to tell them?
  - Who might be harmed or misrepresented by this portrayal?
- Respectful storytelling expands empathy; careless storytelling can reinforce stereotypes or trauma.

Sources: (Haynes, 2021; Khan, 2021; LeClerc, 2024)

Cultural sensitivity and responsibility

- **Cultural sensitivity** is not censorship — it is awareness.
- Portray groups and traditions accurately, based on research and listening.
- Avoid:
  - Using cultural elements as exotic decoration (“appropriation”).
  - Reducing people to symbols or stereotypes.
  - Borrowing trauma without accountability.
- When writing outside your own experience:
  - Consult people from that background.
  - Be open to feedback and revision.
  - Prioritise authenticity over authority.
- Ethical fiction balances creative freedom with empathy and respect for real lives.

Sources: (Khan, 2021; Maguire, 2009; O’Callaghan, 2025)

I am not an expert here

Roadmap

Auditing sources for FAIR and CARE issues

- Next, we'll spend about **30 minutes** discussing and practising what we've learned.
- Two connected parts:
  1. **Discussion of last week's annotated bibliography.** Share one source you found insightful or problematic.
  2. **Source audit activity.** Choose one of your sources and check:
     - Is it **Findable, Accessible, Interoperable, Reusable**?
     - Does it respect **Collective Benefit, Authority to Control, Responsibility, Ethics**?
- Work individually or in pairs; we'll share highlights at the end.

Our goal: to read sources not only for what they say, but for how responsibly they were created.

Your first-draft paper: fitting the scientific style

- The upcoming assignment is your **first draft of a short academic paper.**
- Use the conventions of scientific writing:
  - Clear research question and motivation.
  - Evidence-based argument supported by credible sources.
  - Neutral tone; cautious interpretation; accurate citations.
- Your annotated bibliography and Wikipedia writing experience both help:
  - Bibliography → evaluating sources and citation style.
  - Wikipedia → writing neutrally and clearly for others.

I want to see how well you can use your sources to answer the question, as accurately as possible.

- Next week, you will read each others papers and comment on them.

Acknowledgements

- OpenAI (2025) was used to format the references, and generate a first draft of the slides

Roadmap

References I

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. https://doi.org/10.1145/3442188.3445922

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J., Anderson, J., & Hudson, M. (2020). *The care principles for indigenous data governance* [Global Indigenous Data Alliance]. https://www.gida-global.org/care

Carroll, S. R., Hudson, M., Martinez, A., Kukutai, M. C., & Rodriguez-Lonebear, D. (2020).The care principles for indigenous data governance. *Data Science Journal*, *19*, 1–12. https://doi.org/10.5334/dsj-2020-043

Haynes, P. (2021).The ethics and aesthetics of intertextual writing: Cultural appropriation and minor literature. *British Journal of Aesthetics*, *61*(3), 291–312. https://doi.org/10.1093/aesthj/ayab001

References II

Khan, V. (2021). *A guide to writing cultural diversity in fiction*. Hachette UK. https://www.hachette.co.uk/wp-content/uploads/2021/08/Turning-the-Page-A-Guide-to-Writing-Cultural-Diversity-in-Fiction.pdf

LeClerc, T. (2024).Writing sensitively: Defining sensitivity in the literary sphere. *Journal of Publishing and Society*, *21*(3), 288–298. https://doi.org/10.1080/14790726.2024.2325996

Maguire, E. (2009, September). *Forgive me, forgive me: The ethics of using other people's lives in fiction* [Accessed 2025-11-08]. https://meanjin.com.au/essays/forgive-me-forgive-me-the-ethics-of-using-other-peoples-lives-in-fiction/

O'Callaghan, L. (2025). *Cultural sensitivity: Really useful links*. https://www.writing.ie/resources/cultural-sensitivity-lucy-ocallaghan/

OpenAI. (2025). *Chatgpt (mar 14, 2025 version)* (Large language model). Retrieved October 14, 2025, from https://chat.openai.com/

References III

Smith, J. R., Chou, L. N., & Ghosh, R. (2023). Decorait: A decentralized consent registry for ai training data. https://arxiv.org/abs/2309.14400

Whittaker, M., Crawford, K., & Dobbe, R. (2023).Data extraction, colonialism, and the future of artificial intelligence. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231154690

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016).The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(160018). https://doi.org/10.1038/sdata.2016.18