



DIGITAL TRUST EXPERT GROUP REPORT

2022



CONTENTS

Executive Summary	2
Change Log DTL Criteria Catalog	3
How to use this report	4
Proposals and guidance about existing	4
Label criteria	4
<i>Security</i>	<i>4</i>
Criteria 4	4
Criteria 5	5
Criteria 12	5
<i>Data Protection</i>	<i>6</i>
Criteria 17	6
Criteria 20	6
<i>Fair User Interaction</i>	<i>7</i>
Reflections	9
<i>The Digital Trust Label and B2B2C</i>	<i>9</i>
<i>Reflection on new categories</i>	<i>10</i>
Brainstorm	11
Deep dive	11
Additional research	15
Conclusion and proposals	17
<i>Large language models and DTL</i>	<i>19</i>

DIGITAL TRUST





Executive Summary

The Digital Trust Expert Group (formerly known as Label Expert Committee or LEC) has been actively engaged in the continuous improvement and development of the Digital Trust Label (DTL) to ensure its relevance and effectiveness. To achieve this, the Group has held monthly calls to discuss various aspects of the DTL and its criteria, focusing on the following key topics:

1. **Comments and suggestions for improvement on existing criteria**
2. **Research and proposals on B2B2C (Business to Business to Consumer) relationships and their impact on the DTL**
3. **Proposals and reflections on possible new criteria to strengthen the DTL's comprehensiveness.**

To address these topics more efficiently, the Digital Trust Expert Group has divided into working groups based on the existing criteria dimensions. Each working group has taken on the responsibility of addressing direct requests from the SDI, working on the criteria with a focus on the above-mentioned priorities. Furthermore, the working groups have explored potential new criteria and presented their proposals in this report.

Dimension Security. Alongside general reflections on the Security dimension, the Digital Trust Expert Group suggests a slight modification of the wording for criteria 5 and 12. Notably, the Group propose a split for criterion 12 to address different aspects more effectively.

Data Protection. In terms of the Data Protection dimension, the Group do not propose any changes. The reasons for this decision are explained in the corresponding section of the report. Inquiries from SGS and responses from the Digital Trust Expert Group are documented and commented on in the relevant section. The group reiterates that the Digital Trust Label is compatible also with the revised version of the Swiss Data Protection legislation.

Reliability. For the Reliability dimension, the Digital Trust Expert Group has no specific comments or suggestions for expansion.

Fair user interaction. Within the Fair User Interaction dimension, the Digital Trust Expert Group has primarily focused on the metrics of Transparency and Explainability. Through discussions and research, the Group has developed a foundational understanding of how these metrics can be described in general and which subject areas will be in focus for this year. With the rapid development of Large Language Models, the Expert Group will pay particular attention to these two metrics to ensure the DTL stays relevant and up-to-date regarding technological advancements.

Change Log DTL Criteria Catalog

Current criteria	Proposed change (in bold)
<p>12: Critical security vulnerabilities shall be communicated to relevant authorities within 72 hours if not corrected, and the impacted users shall be timely and adequately informed. Personal data breaches shall be communicated to relevant authorities and impacted data subjects within 72 hours.</p>	<p>12: Critical security vulnerabilities shall be communicated to relevant authorities within 72 hours if not corrected, and the impacted users shall be timely and adequately informed if there is an update to be installed</p>
<p>new</p>	<p>13: Personal data breaches that create high risks for users shall be communicated to relevant authorities and impacted data subjects within 72 hours.</p>
<p>13: The user shall be informed about the purpose of the processing and the legal basis for processing of their personal data in clear and plain language. Where there is more than one purpose/legal basis, they need to be listed separately in a way that the user is able to easily distinguish between one purpose/legal basis and another.</p>	<p>14: The user shall be informed about the purpose of the processing and / or the legal basis for processing of their personal data in clear and plain language. Where there is more than one purpose and /or legal basis, they need to be listed separately in a way that the user is able to easily distinguish between one purpose and / or legal basis and another.</p>
<p>14: Where user consent is sought for the processing of personal data, such consent shall be expressly collected from the user for each of the purposes and legal basis listed by the service provider and obtained separately from the terms and conditions of use of the services.</p>	<p>15: Where user consent is sought for the processing of personal data, such consent shall be expressly collected from the user for each of the purposes and / or legal basis listed by the service provider and obtained separately from the terms and conditions of use of the services.</p>
<p>16: The user shall be provided with a separate, easy, and accessible way of withdrawing consent.</p>	<p>17: The user shall be provided with a separate, easy, and accessible way of the right to object.</p>



How to use this report

This report summarizes the discussions held among the Digital Trust Expert Group over the course of the last year (June 2022 – May 2023). The purpose of the document is to propose changes to the criteria listed in the DTL Criteria Catalog (“Proposals”) as well as to provide guidance to the auditors when it comes to the interpretation of the criteria (“Guidance”). The report follows the structure of the DTL dimensions and goes through all the criteria where the experts have discussed comments. A section on “Reflections” at the end shows which questions will be debated going forward and need more attention as well as thoughts by the experts on emerging issues.

Proposals and guidance about existing label criteria

Security

Criteria 4

Secure user authentication is an essential aspect of maintaining the safety and integrity of digital services. As security threats evolve, it is crucial to explore and implement diverse authentication methods that go beyond traditional password-based systems. Employing a risk-based approach, tailored to the specific requirements and sensitivities of the service in question, can help enhance security without sacrificing user experience. In addition to password-based authentication, token-based systems, and single sign-on (SSO) solutions can provide more robust security.

Two-factor authentication (2FA) is another important measure to consider when designing secure authentication systems. By requiring users to provide two separate forms of verification – typically, something they know (e.g., a password) and something they possess (e.g., a smartphone) – 2FA makes it more difficult for attackers to gain unauthorized access. However, it is essential to note that while 2FA provides an additional layer of security, it is not a foolproof solution and should be employed as part of a comprehensive security strategy.

Moreover, secure authentication should not only be a priority for external users but also for internal systems and personnel. Default authentication settings are often weak and easily exploitable, making it critical for organizations to establish strong internal authentication protocols. These may include employing risk-based access controls, ensuring the principle of least privilege is followed, and implementing proper security training for employees.

It is important to recognize that not all privacy-enhancing technologies possess the same level of sophistication. Some solutions may offer robust security features, while others may provide only basic protection. Therefore, when implementing privacy-enhancing technologies, it is crucial to evaluate each solution's efficacy and compatibility with other security measures to create a comprehensive and effective security framework.



Criteria 5

Guidance: 5: Guidance for secure installation, configuration, and updates shall be in place for both internal changes and users, and updated for each release if necessary. Guidance shall be available in a manner that is easy to access and understand. Any major changes shall lead to a communication to the users in an easy-to-understand format.

Criteria 12

Part 1: Addressing Critical Security Vulnerabilities

Organizations should diligently address critical security vulnerabilities, taking into consideration the guidance provided by the Standardization Guidance System (SGS). It is concerning that many companies have not yet adequately tackled these vulnerabilities, leaving their systems and user data exposed to potential breaches. To mitigate these risks, organizations must prioritize identifying and rectifying critical security vulnerabilities as soon as possible.

Upon the discovery of a critical security vulnerability, organizations are required to notify relevant authorities within 72 hours if the issue remains uncorrected. This timely communication allows authorities to monitor and assess the potential impact of the vulnerability on a wider scale. Furthermore, impacted users must be promptly and adequately informed if there is an update or patch to be installed to resolve the vulnerability. This ensures that users can take necessary actions to protect their data and devices from potential threats.

Part 2: Reporting Personal Data Breaches


In the event of a personal data breach that poses high risks for users, organizations must adhere to strict reporting guidelines. Relevant authorities and impacted data subjects should be notified within a 72-hour window following the detection of the breach. This prompt communication enables users to take appropriate measures to safeguard their personal information and minimize potential harm.

Organizations should also implement effective incident response plans and protocols to manage personal data breaches. This includes conducting thorough investigations to determine the scope and severity of the breach, taking appropriate remediation actions, and identifying measures to prevent similar breaches from occurring in the future.

In conclusion, it is imperative for organizations to address critical security vulnerabilities promptly, with guidance from SGS, and to report personal data breaches that present high risks to users. By adhering to these guidelines and prioritizing security, organizations can better protect user data and maintain the trust and confidence of their customers.

Proposals

Split this criterion into two parts. Change the first part regarding critical security vulnerabilities based on the guidance of SGS. Decide what to do about the fact that most companies have not yet addressed critical security vulnerabilities. Hence, the Expert Group proposes the following:



12: Critical security vulnerabilities shall be communicated to relevant authorities within 72 hours if not corrected, and the impacted users shall be timely and adequately informed **if there is an update to be installed.**

13: Personal data breaches **that create high risks for users** shall be communicated to relevant authorities and impacted data subjects within 72 hours.

Data Protection

Data protection is currently in an evolutionary state. Especially in Switzerland the revised Federal Act on Data Protection (FADP) will come into force on September 1st 2023. The revised FADP is based on the EU's counterpart the GDPR. Besides many similarities that make the two acts more or less compatible, there are some differences in the details. Differences include, for example, deadlines for the reporting of breaches, the way fines are enforced, or default principles for processing personal data.

Given the evolutionary state, the fact that from the auditors' side there seems to be no urgent requirements to change the existing data protection criteria, and last but not least the fact that the existing criteria are rather new, **we do not see any necessity to change or extend the existing set of criteria.**

Criteria 17


Text in the DTL: "The user shall be informed of the definite time period for which the personal data will be stored. If that is not possible, the user shall be informed of the criteria and reasons used to determine the indefinite period, and a regular timeframe for which a review will be undertaken."

Comment SGS: "The user shall be informed of the definite time period for which the personal data will be stored. Consider adding what definite time period is considered appropriate. For instance, cookie retention period greater than 1 year are considered inappropriate Additional information: <https://www.mofo.com/resources/insights/220504-cookie-consent-requirements>."

Guidance Digital Trust Expert Group: We believe that often it does not make sense to mention a definite time period, since this might vary from data object to data object (e.g., cookie vs. payment information). However, the last sentence in the criteria does not make much sense, since the user (typically) does not get informed about a review. We recommend omitting the last sentence.

Criteria 20

Text in the DTL: "The service provider shall ensure that the user can access their data. Any requests for access need to be acceded to within 30 days. Together with a copy of the personal data, a user is to be provided with names of third parties with whom such personal data has been shared, together with the legal basis under which such data is being held."



Comment SGS: "Typically the process does not include the sharing of the third parties within the Data Subject Access Request (DSAR)"

Guidance Digital Trust Expert Group: "We believe it doesn't need to be a detailed list of data recipients, but could be summarized in the sense that the user knows about the fact that data is shared with third parties. Last but not least, we want to emphasize, that the label may be stronger than the legal requirements (the legal requirements have to be met anyway)."

Proposals Regarding Data Protection Criteria

Criteria 13: «The user shall be informed about the purpose of the processing and / or the legal basis for processing of their personal data in clear and plain language. Where there is more than one purpose and /or legal basis, they need to be listed separately in a way that the user is able to easily distinguish between one purpose and / or legal basis and another.»

Criteria 14 and 16 (in view of criteria 15) have to be considered too:

Criteria 14: «Where user consent is sought for the processing of personal data, such consent shall be expressly collected from the user for each of the purposes and / or legal basis listed by the service provider and obtained separately from the terms and conditions of use of the services.»

Criteria 16: «The user shall be provided with a separate, easy, and accessible way of the right to object.»


Guidance: From what has previously been discussed the Expert Group concludes that **a Swiss-based company offering a digital service to Swiss customers should be awarded the DTL if they follow the revised Swiss data protection legislation.**

Fair User Interaction

The best matching category to include explainability and transparency related metrics seems to be "Fair user interaction" - it is also the only one that explicitly mentions the user. We are aware that in the AI ethics community, the fairness principle is often limited to questions of (non-)discrimination. However, the main audience of the DTL is the public at large, so we consider we can extend "fair user interaction" to additional concerns.

As such, we propose to include the following new topics in the "fair user interaction" dimension:

Explainability and transparency are crucial aspects of ensuring fair user interaction with AI systems. These concepts encompass the ability of AI systems to clearly communicate their decision-making process, rationale, and potential biases, as well as the overall accessibility of their inner workings to users and stakeholders.



1. Explainability: refers to the ability of an AI system to provide human-understandable explanations for its decisions, predictions, and proposals. Explainable AI models can be interpreted by users, helping them understand why and how the system arrived at a particular output. This understanding allows users to make informed decisions and assess whether the AI system is functioning as intended.

Some key aspects of explainability include:

- a) Local explainability: Offering explanations for individual predictions or decisions made by the system.
- b) Global explainability: Providing a general understanding of how the system makes decisions across a wide range of inputs.

2. Transparency: refers to the openness and accessibility of an AI system's design, decision-making process, and data handling practices. This includes providing information about the training data, algorithms, and other factors that influence the AI's output. Transparent AI systems enable users to examine and assess the fairness, accuracy, and reliability of the system.

Some key aspects of transparency include:

- a) Algorithmic transparency: Disclosing the algorithms and techniques used in the AI system to ensure users can understand and scrutinize the decision-making process.
- b) Data transparency: Sharing information about the training data, such as its sources, quality, and potential biases, to allow users to assess the AI system's fairness and reliability.

Promoting explainability and transparency can lead to a variety of benefits, such as:

1. Trust and confidence: When users understand how AI systems work and the rationale behind their decisions, they are more likely to trust the system and feel confident using it.
2. Accountability: Explainable and transparent AI systems enable users to hold developers and operators accountable for the system's performance, ethical considerations, and potential biases.
3. Informed decision-making: Users can make better-informed decisions when they understand the AI system's decision-making process, increasing the likelihood of fair outcomes.
4. Feedback and improvement: Explainability and transparency allow users to provide feedback on AI system performance, which can drive improvements in system design and operation, ultimately leading to more fair and ethical AI systems.

Explainability and transparency are essential for fair user interaction with AI systems. By fostering an environment where users can understand and assess the decision-making process and the factors influencing AI outputs, developers can build trust, promote accountability, and ensure ethical AI use.




Reflections

The Digital Trust Label and B2B2C

We list some reflections on why it is a complex task to propagate the label in a B2B2C environment.

1. Complexity of relationships: In a B2B2C setting, the relationships between businesses and consumers are more complex, as multiple parties are involved. Ensuring that all parties adhere to security, data protection, and fair user interaction standards can be difficult.
2. Varying standards and regulations: Different companies may operate under different jurisdictions and follow different data protection and security standards. Ensuring that all parties involved comply with the same standards can be challenging.
3. Supply chain risks: There may be multiple layers of subcontractors and suppliers in a B2B2C setting, each with their own security and data protection practices. Assessing and managing the risks associated with each party can be difficult and time-consuming.
4. Limited visibility and control: The company owning the Trust Label might not have full visibility into the practices of its clients and their partners. This lack of control makes it harder to assess and guarantee compliance with security and data protection standards.
5. Reputation risk: If the client company fails to uphold the standards set by the Trust Label, it could harm the reputation of the company owning the trust label, as they are indirectly associated with the client's practices.
6. Resource-intensive process: The process of assessing and monitoring a client company's security, data protection, and user interaction practices can be resource-intensive, requiring significant time and effort to ensure compliance.
7. Evolving threats and technologies: Security threats and technology landscapes are constantly changing. Ensuring that a client company stays up-to-date with the latest security measures and adapts to new threats can be difficult.
8. Scalability: As the number of clients and partners increases, it becomes more challenging for the company owning the Trust Label to ensure that all parties involved maintain the required security and data protection standards.
9. Conflicting interests: There might be conflicting interests between the parties involved in a B2B2C setting, which could make it difficult to enforce certain security or data protection practices.
10. Legal and contractual issues: Ensuring that all legal and contractual obligations related to security, data protection, and user interaction are met by all parties involved can be complicated, especially when dealing with multiple jurisdictions and legal systems.

We identified challenges where organizations build their products based on pre-trained large language models.




In the case of a company building products based on pre-trained large language models, there are additional challenges when it comes to awarding a Digital Trust Label regarding security, data protection, and fair user interaction. Some of these challenges include:

1. **Data privacy concerns:** Large language models are trained on vast amounts of data, often from various sources. Ensuring that the training data used complies with privacy regulations and that no personally identifiable information (PII) is included can be difficult.
2. **Model transparency:** Understanding the inner workings of large language models can be challenging, making it difficult to assess how secure, fair, and privacy-preserving they are. Ensuring transparency in model development and deployment is crucial for maintaining trust.
3. **Bias and fairness:** Large language models can inadvertently learn and propagate biases present in the training data. Ensuring that the models are unbiased and fair in their outputs and interactions with users is essential for maintaining trust.
4. **Output control:** Large language models can sometimes generate unexpected or inappropriate outputs. Ensuring that the generated content adheres to ethical standards and complies with relevant regulations can be challenging.
5. **Intellectual property:** Since language models are trained on vast amounts of data, it's possible for them to inadvertently generate content that infringes on intellectual property rights. Ensuring compliance with IP regulations can be difficult.
6. **Security vulnerabilities:** Large language models can be susceptible to adversarial attacks or other security vulnerabilities. Ensuring that the models are secure and robust against potential attacks is crucial for maintaining trust.
7. **Access control:** Products based on large language models may require strict access control to prevent unauthorized use or misuse. Ensuring that proper access control mechanisms are in place and are maintained can be challenging.
8. **Ongoing monitoring:** Given the rapidly evolving nature of AI technologies, maintaining trust in large language models requires continuous monitoring and updating to address new security, data protection, and fairness concerns as they arise.
9. **Model interpretability:** Large language models can be complex and difficult to interpret, making it harder to provide explanations for their outputs or decisions. Ensuring that the models can be easily understood and explained is important for maintaining trust.
10. **Legal and ethical considerations:** As AI technologies continue to advance, there may be new legal and ethical concerns that arise. Ensuring that the company building products based on large language models stays up-to-date with and adheres to relevant regulations and ethical guidelines is crucial for maintaining trust.

Reflection on new categories

In addition to the strengthening of existing label criteria and the formulation of new criteria within the four DTL categories, the Digital Trust Expert Group considered the addition of new categories and criteria within those categories. Might there be topics not yet covered by the Digital Trust Label that should be taken into account for it to remain a credible proxy for digital trust?



The Digital Trust Expert Group followed a four-step process to explore the potential addition of new categories to the DTL catalogue:

1. A brainstorm involving the full Digital Trust Expert Group
2. A deep-dive by a dedicated working group into the topics identified in the brainstorm
3. Additional research
4. Formulation of conclusions and proposals to the SDI board

Brainstorm

Like for the existing DTL categories, the Digital Trust Expert Group used an online Mural to gather and discuss ideas. The brainstorm session resulted in the identification of four potential new categories:

- 1) human rights
- 2) sustainability
- 3) transparency & accountability
- 4) usability

It was decided to form a dedicated working group to conduct a deep dive on these topics. The working group benefited from contributions by Nikki Boehler (sustainability), Sophia Ding (transparency and accountability), Maximilian Groth (usability), Rodolphe Koller (transparency and accountability), Diego Kuonen (transparency and accountability), Charlotte van Ooijen (group lead and sustainability), Leila Topic (human rights).

Deep dive

The members agreed to focus on three aspects for the deep-dive, presented in order of priority:

1. **Motivation** to include the new category: Considering the latest academic research, legal standards, public debate and business developments, why is it relevant to consider this aspect when assessing the trustworthiness of digital services?
2. **Relation** with existing categories: Do the new categories have a link with any existing category, potentially leading to the adoption of formulated criteria by the other working groups?
3. What **tentative criteria** could be formulated for every potential new category?

Human Rights

Motivation

- Technology systems have increasingly come under criticism for creating or exacerbating negative impacts on a range of human rights.
- The proposed criteria will ensure that DTL assesses organizations' policies and procedures for addressing risks to human rights linked to their technology systems and encourages them to adopt approaches that are aligned with the business responsibility to respect human rights, as laid out by the UN Guiding Principles on Business and Human Rights (UNGPs).
- This includes technology itself, business model, and Go-To-Market (GTM)



Tentative criteria

Does the organization assess human rights risks associated with their system?

- Does the organization take steps to mitigate human rights risks? (e.g. built-in technological safeguards, internal escalations, regular testing/modifying of its technologies, contractual and policy safeguards, and training for customers/users.
- Does the organization take steps to provide access to remedy for individuals that are exposed to the human rights risks associated with the organization's system?
- Does the organization consider the risks to human rights associated with high-risk customers or users?


Sustainability

Motivation

There are several reasons to consider including criteria related to sustainability in the DTL catalogue.

1. There is a moral and legal imperative for all inhabitants on Earth, and notably governments and big corporations, to combat and prepare for climate change. The Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), the world's leading authority on climate science presents the increased evidence of climate change and the "unequivocal" human influence on the warming of the atmosphere, ocean and land. In July 2022, the UN General Assembly adopted a historic resolution, declaring access to a clean, healthy and sustainable environment, a universal human right, thereby moving the sustainability question out of the realm of discretionary policy and into that of legal obligation. The resolution calls upon States, international organizations, and businesses to scale up efforts to ensure a healthy environment for all. The resolution was previously adopted by the UN Human Rights Council in October 2021.

The European Commission has included a chapter on sustainability into the European Declaration on Digital Rights and Principles for the Digital Decade: "To avoid significant harm to the environment, and to promote a circular economy, digital products and services should be designed, produced, used, disposed of and recycled in a way that minimises their negative environmental and social impact. Everyone should have access to accurate, easy-to-understand information on the environmental impact and energy consumption of digital products and services, allowing them to make responsible choices."



2. Research has shown that digital technologies and data can have both a positive and negative impact on the sustainable development goals. Digital services

a) help society adapt to the consequences of climate change, eg by better predicting extreme weather events and alerting the population in a timely way (See DECI DO project)

b) help mitigate climate change, eg through smart grids and energy consumer awareness

c) leave a significant carbon imprint through data processing power. The information and communication industry's contribution to global greenhouse gas emissions was estimated to fall between 3.0 and 3.6% of total emissions in 2020, with the potential to exceed 14% of the 2016 net emissions in 2040 (Belkhir and Elmeligy, 2018).

3. The public more and more expects service providers, including digital service providers, to address the sustainability question. Several providers have publicly announced undertaking efforts to become carbon-neutral (e.g. EY and Microsoft). Some dedicate a section or page on their website to their company's contribution. (illustrative best practices to be added). Others have included a service module in their digital transactions, mentioning the service's environmental impact and offering a way to offset it (NB the latter in some cases turns out to be greenwashing rather than real compensation). (NB research to be added on consumers expectations regarding sustainability. eg eurostat)

4. Thus, both from a moral, legal and social point of view, digital service providers should pay attention to the sustainability of their services. Ideally, service providers should 1) assess and 2) publish in a transparent manner the environmental impact (positive and negative) of their service and 3) demonstrate efforts to reduce the negative impact to a minimum. However, the Digital Trust Expert Group is aware that such assessment may be a complicated effort, especially for smaller organisations. Therefore, service providers must demonstrate that they have started making concerted efforts in the right direction regarding both their sustainability policy and environmental transparency

5. Sustainability can be considered relevant for the DTL because "trust in the digital world [...] encompasses social and ethical responsibility." (SDI Digital Trust Whitepaper, 2022)



Transparency and accountability

Motivation

- Transparency, Accountability and Auditability are currently partially covered within other categories, e.g., criterion 13, 17, 22, 24, 27, 28, 29, 34.
- Does it make sense to keep the transparency-related criteria within the other groups or create a horizontal category?
- Create a new group for accountability (currently part of reliability) or rename it to reliability & accountability?
- NB “Transparency is key in building Digital Trust [...] presenting relevant information for informed decision-making in a clear fashion.” (SDI Digital Trust Whitepaper, 2022)
- Link to other label categories

Usability

Motivation

The motivation to include usability is based on the following three components:

1. Label's perspective: Usability is key for the Label to remain up to date and relevant.
2. End user's perspective: if the services are hard to use or made complicated, because of the Label's requirements, he/she/they might lose interest.
3. Customer's perspective: To make it as effortless by providing usability guidelines.

One can use the findings of [eGovernment benchmark usability pilots. A study from the Lisbon Council and Public.Digital defines and pilots eGovernment usability indicators to ensure they are up to date](#) provides an initial orientation:

1. Does the organisation use a clear language?
2. Is the usability consistent and ease of use?
3. Do speed and performance comply with modern standards?
4. Is help and support available within a reasonable effort?



Additional research

The working group members presented the results of the in-depth exploration to the other Digital Trust Expert Group members to gauge the relevance of the new categories and decide on next steps. While acknowledging the social significance of the human rights and sustainability categories, the Digital Trust Expert Group did not demonstrate sufficient support to develop criteria on these topics for the DTL catalogue. However, it was agreed that additional explorations were justified to support the formulation of proposals to the SDI board. On transparency and usability, it became clear that the work undertaken on both topics merits integration into existing label categories. The findings from an additional round of research and deliberations point to more precise implications for the DTL catalogue and process.

Human Rights

Several international standards are under development impacting organisations' requirements regarding reporting on human rights.

In November 2022, the EU Parliament adopted The Corporate Sustainability Reporting Directive (CSRD) which will require that companies take a "double materiality" approach to disclosure seriously (i.e. impact on the world beyond the company's financial value). CSRD introduces more detailed reporting requirements on companies' impact on the environment, human rights and social standards... To ensure companies are providing reliable information, they will be subject to independent auditing and certification.

First set of standards are under development and will be adopted by June 30th 2023.

The standards are required to specify the information that should be disclosed regarding 4 categories - Environment, Social, Governance, and Cross-Cutting standards. Social – covers (1) Own Workforce, (2) Workers in the value chain, (3) Affected communities, and (4) Consumers and end users. Including respect for human rights, fundamental freedoms, democratic principles and standards established in the International Bill of Human Rights and other core UN human rights conventions.

Sustainability

Further research revealed there is a [Sustainable IT label](#), "[Le label NR](#)", which was equally identified by the SDI as part of an [inventory of international labels](#). The research undertaken by the Digital Trust Expert Group into the NR label specifically complements and updates the insights of the SDI inventory. Le label NR was initiated in 2019 in France by the French Institute for Sustainable IT and has since then been deployed in Belgium and Switzerland. The discovery of this label raises the question whether it would be redundant for the DTL to include criteria on this topic if these are already covered by a sister label. To learn more about the Label NR's mission and methodology, and the potential overlap, complementarity and synergies with the DTL, on 8 December 2022 Nikki Boehler and Charlotte van Ooijen had an exchange with several representatives:

- Rémy Marrone, projects director at the French, Belgian and Swiss [Institutes for Sustainable IT](#), the associations responsible for the development of the label;
- Jocelyn Oppenlander, co-founder of the Label NR in Switzerland and secretary of ISIT-CH;
- René Masson, project manager



Based on that meeting and desk research it has become clear that:

- The Label NR has been founded and developed in France, as a mandate by WWF and with state support. It has been deployed in other French speaking countries as well: Belgium and Switzerland. They have a very clear governance structure. The management of the NR label is carried out by Agence LUCIE in France, the Institute for responsible IT (Institut du Numérique Responsable-[INR](#)) in Switzerland and the Belgian Institute for a Sustainable IT in Belgium.
- The main focus is on reducing the ecological footprint of digital technologies. They do not accredit services, but all types of organisations: public sector, large corporations, SMEs and administration. According to the team, the biggest demand comes from the public sector. While the NR label has mainly accredited French organisations, it is also in the process of accreditation of three Swiss organisations: (the city of Lausanne, Canton of Geneva and Darest Informatique in Geneva) and accredited one so far (Services Industriels de Genève).
- The label, available at two levels, is based on: a self-assessment framework, management by Agence LUCIE of the labelling process, a commitment from the candidate organisation to contribute to the operation of the system, an audit by Agence LUCIE and for level 2 an audit review by reference organisations (SGS, Bureau Veritas or Baker Tilly STREGO). Finally, the label is awarded by a labelling committee of Green IT experts.
- Their ambition is to expand the label NR across Europe, for which early 2023 a European institute, the ISIT Institute, will be launched to lead the effort. They aim to set up teams in every country that they are active in and already have an operational team in Switzerland.
- Currently, the NR label catalogue already includes criteria that go beyond the topic of sustainability, such as responsible data management. The label NR team expressed a clear ambition to expand the label even further from sustainable to responsible IT, as the French label name already implies. Therefore, they are thinking about including ethical and security aspects in their label.

In light of the overlap between the DTL and the label NR, both in terms of overall mission, label criteria and international ambitions, it is advisable for both labels to at the least coordinate and at the best actively look for synergies and collaboration. The Digital Trust Expert Group wishes to underline that a label is a means to an end, namely fostering trustworthy and responsible digital services, and not an end in itself. As such, competition between labels should be avoided. The DTL and other labels should take care that the ambitions for expansion never stand in the way of achieving the overall shared goal.

Transparency and accountability

On transparency and accountability, criteria need to be formulated and harmonised across the other existing categories; Three group members worked on horizontal process-related characteristics that may be applied to all transparency-related criteria across categories.

Conclusion and proposals

The aforementioned exploration has led the Digital Trust Expert Group to the conclusion that **at this time it is not advisable to extend the label catalogue with new categories**. However, the undertaken work does reveal the relevance of all four topics for discussions on digital trust and therefore the context in which the DTL operates. In order to uphold the legitimacy and credibility of the DTL's content and process, the Digital Trust Expert Group considers it of utmost importance for the SDI board to take action on all four topics. The table provides an overview of the conclusions per explored topic and the proposals to the SDI board.

Topic	Conclusion	Proposals
Human rights	<p>Technology products have increasingly come under criticism for creating or exacerbating negative impacts on a range of human rights. There is no trust in the digital age without actually knowing the digital systems respect and protect your rights.</p> <p>The UN Guiding Principles on Business and Human Rights (UNGPs) are the authoritative global standard concerning business impacts on people, including those affected by the use of a company's products and services. Under the UNGPs, companies are expected to conduct human rights due diligence across all business activities and relationships. DTL needs to align with UNGPs.</p>	<ul style="list-style-type: none"> • No need to include criteria on human rights. • Explicitly communicate alignment with / support for human rights in DTL communications. • Point to credible external resources such as the work of OHCHR's B-Tech initiative.
Sustainability	<p>Sustainability is undeniably a relevant criterion for responsible digital services. However, its impact on digital trust is not sufficiently evident at this time. While the formulation of new criteria is therefore not warranted for the current revision of the DTL, it is imperative for the SDI to otherwise acknowledge and promote the importance of sustainability in digital service delivery.</p>	<ul style="list-style-type: none"> • No need to include criteria on sustainability • Explicitly underline the importance of sustainability in communications accompanying the DTL while clearly stating that the DTL does not include criteria on sustainability • Strong advice to further explore the relation between digital trust and sustainability • Explore synergies with sustainability-focused labels, especially the label NR • Closely follow the public and political debate on sustainable digital services and the relation with digital trust to adapt the DTL as necessary



Transparency		<ul style="list-style-type: none">• Need to harmonise transparency-related criteria in existing categories• Add criteria to fair user interaction (see detailed conclusion there)
Usability		<ul style="list-style-type: none">• Add criteria to fair user interaction/reliability (see detailed conclusion there)



Large language models and DTL

The challenges with Large Language Models are intertwined with the Digital Trust Label. Every dimension is affected by this. In this year (2023), we will increasingly address the technological developments and analyze their impact on the different dimensions of the Trust Label.

In particular, we will investigate the following subject areas:

1. Bias and fairness: large language models are trained on vast amounts of data from the internet, which may include biased, controversial, or offensive content. As a result, these models can inadvertently learn and perpetuate these biases, **leading to unfair or discriminatory outputs**. This can negatively impact digital trust as users may question the fairness and ethical use of such models.
2. Explainability and transparency: As these models grow in size and complexity, it becomes **increasingly difficult to understand their inner workings and decision-making processes**. This lack of explainability and transparency can hinder users' ability to trust the AI system and evaluate its reliability, as they cannot easily assess the rationale behind the model's outputs.
3. Misuse and malicious applications: large language models can generate highly realistic content, which can be exploited for malicious purposes, such as spreading misinformation, deepfake creation, or phishing attacks. **These negative uses can significantly undermine digital trust**, as users may become wary of interacting with AI-generated content or using AI systems in general.
4. Data privacy and security: Training large language models requires massive amounts of data, **which may include sensitive or personally identifiable information**. Ensuring the privacy and security of this data is crucial for maintaining digital trust. Failing to protect user data or using it without proper consent can lead to mistrust and skepticism.
5. Accountability and responsibility: Determining accountability and responsibility for the outputs and actions of large language models can be complex, as the line between the AI system, its developers, and its users can be blurred. **Establishing clear guidelines and mechanisms for accountability** is essential for building digital trust.
6. Filter bubbles and echo chambers: large language models can inadvertently reinforce users' existing beliefs and preferences by generating content that aligns with their interests. This can lead to filter bubbles and echo chambers, where users are only exposed to information that confirms their existing beliefs. **This phenomenon can further erode digital trust** by promoting misinformation and polarizing perspectives.

SWISS DIGITAL INITIATIVE

CONTACT

Swiss Digital Initiative
Campus Biotech
Chemin des Mines 9
1202 Geneva

info@sdi-foundation.org

swiss-digital-initiative.org
digitaltrust-label.swiss