



HEALTH
DATA HUB



BEACON

**Ce document a été réalisé en partenariat avec Veltys
et avec la contribution de l'INRAE.**



Description et type de standard : Beacon est à la fois un protocole de requête, une interface de programmation (API) et un modèle de données qui permet **le partage et la découverte de données génomiques et biomédicales conservées dans des bases de données dispersées** pour les chercheurs qui souhaitent identifier la présence de variations génétiques. Le standard Beacon a été créé en 2013 par la Global Alliance for Genomics and Health (GA4GH).

Domaines d'application : Recherche **génomique** et **clinique**.

Maturité / Utilisation : Beacon est utilisé par plus de **40 organisations internationales**, sur plus de **200 bases de données** concernant plus de **100 000 individus**. En particulier, l'infrastructure de bio-informatique européenne **ELIXIR** a constitué un **réseau de Beacons européens** (voir l'élément « Adoption du standard » en partie 4. Valorisation).

1. Général



PRÉSENTATION

- **Pays d'origine** : international¹.
- **Consortium d'origine** : Global Alliance for Genomics and Health (GA4GH) créée en 2013 pour fonder un système fédéré de partage de données génomiques et d'informations cliniques associées².
- **Type de standard** : format de transfert.

DESCRIPTION

Beacon est à la fois un **protocole de requêtage**, une **interface programmatique de transfert de données (API)** - aussi apparenté à un web service - et un **modèle de données** qui permet le **partage** et la **découverte de données génomiques et biomédicales** dans des bases de données dispersées pour les chercheurs qui souhaitent identifier la présence de variations génétiques³. Le concept de Beacon a été présenté dès la fondation de la Global Alliance for Genomics and Health (GA4GH) en 2013 comme un moyen de connecter les producteurs de données génomiques et les chercheurs souhaitant y accéder. **Le projet Beacon a été lancé en 2014**⁴.

- Le modèle de données Beacon et son implémentation en standard de web service (API) constituent les spécifications « Beacon ». L'implémentation de ces spécifications dans des bases de données individuelles permet la découverte de données⁵ génomiques et phénocliniques.
- Initialement, le protocole Beacon (**v0 et v1**) permettait uniquement des requêtes sur la présence/l'absence d'une mutation génétique spécifique dans un ensemble de données (sur des patients ayant une maladie donnée ou sur la population en général, voir Figure 1 ci-dessous).

1) La Global Alliance for Genomics and Health est un consortium international principalement financé par les institutions suivantes : Canadian Institutes of Health Research, Genome Canada, UK Research and Innovation, National Institute for Health Research, National Human Genome Research Institute, Wellcome.

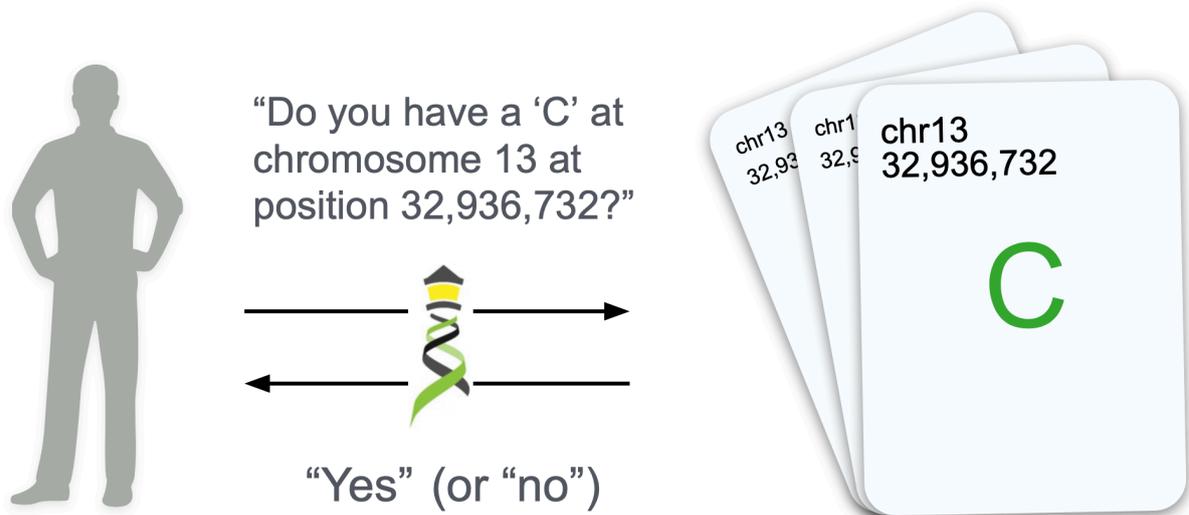
2) Voir l'article de Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). « Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond ». *Human Mutation*, 43, 791– 799. <https://doi.org/10.1002/humu.24369>

3) Voir : [GA4GH Beacon project](#)

4) Voir : [Extensions to the GA4GH Beacon API will enable a more powerful community resource](#)

5) La découverte de données, ou data discovery, est un processus qui consiste à agréger les données et à les visualiser afin de détecter des patterns. Un exemple d'outil de data discovery appliqué à la génomique et développé par Genomics England (société britannique créée et détenue par le Ministère de la Santé et des Affaires sociales du Royaume-Uni pour gérer le projet « 100 000 génomes ») : [Data Discovery - Genomics England Research Environment User Guide](#)

Figure 1 : Exemple schématique d'une requête Beacon jusqu'à la v1



Source : [GA4GH Beacon project](#)

- La nouvelle version du protocole Beacon (**v2**) permet des requêtes plus riches (voir Figure 2)³ :
 - **Ajout de nouvelles options de requêtage pour des métadonnées** biologiques ou techniques en utilisant des filtres (ex : code identifiant une maladie spécifique, sexe, âge⁶).
 - Options pour la collecte des données (ex : si Beacon est utilisé en interne dans un établissement, l'option permet d'indiquer l'identifiant du dossier des patients présentant la mutation d'intérêt).
 - Options pour l'accès aux données : informations utiles au **processus de collectes de données** (ex : personnes à contacter, conditions d'utilisation des données).
 - Annotations sur les **variants** (ex : conclusion des experts/médecins sur la pathogénicité d'une mutation donnée, ou son rôle dans la production d'un phénotype donné).
 - Informations sur les **cohortes**.

6) Pour plus d'informations : [Filters - Beacon v2 Documentation](#)

Figure 2 : Exemple schématique d'une requête Beacon (v2)



Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?



Beacon v2 API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "genomics API".

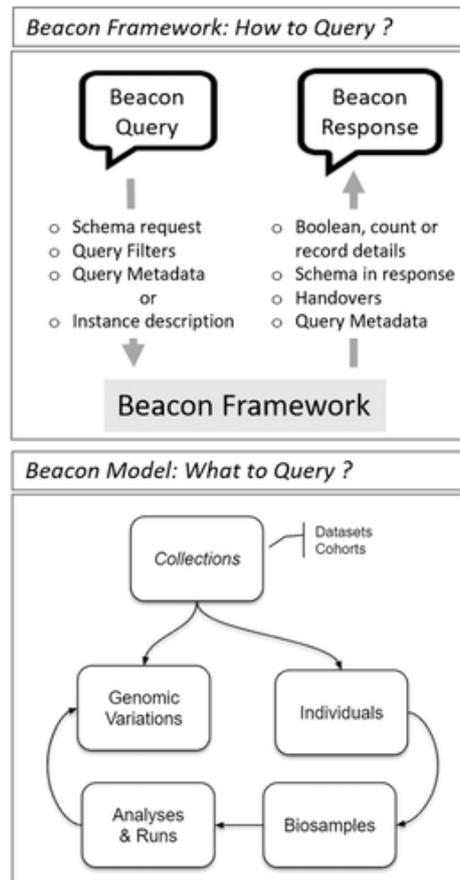
Source : [GA4GH Beacon project](#)

- **La spécification Beacon v2** est composée de deux éléments interdépendants (voir Figure 3 pour une illustration) :
 - **Le Framework Beacon** : décrit la structure globale des requêtes API et des réponses, des paramètres, des composants communs, etc. Dans ce Framework, les options des requêtes sont distinguées des réponses associées. Il existe plusieurs types de réponses possibles (aussi appelées : Beacon « Flavours », voir Figure 3 ci-dessous)⁷ :
 - **Booléen** : Oui / Non.
 - **Dénombrement** : Oui / Non et le nombre de résultats correspondant.
 - **Ensemble de résultats** : Oui / Non, le nombre de résultats correspondant et la décomposition pour chaque collection (base de données ou cohorte, voir l'élément « Complexité du modèle » en partie 3. Technique), et éventuellement des détails sur chaque enregistrement qui correspond à la requête.
 - **Le Modèle Beacon**⁸ : décrit l'ensemble des concepts inclus dans une version Beacon (ex : v2), tels qu'un individu ou un échantillon biologique. Il existe un Modèle par défaut qui contient les entités utiles pour les données génomiques et cliniques, mais ce Modèle n'est pas imposé (voir l'élément « Complexité du modèle » en partie 3. Technique pour la description des 7 entités qui constituent le Modèle par défaut).
 - Une **instance Beacon** est l'implémentation d'un Modèle Beacon qui suit les règles établies par le Framework Beacon.

7) Voir : [Beacon v2 Framework](#)

8) En principe, il est possible d'adopter différents modèles en utilisant le même Framework (ex : « Imaging Beacon », [About the Precision Imaging Beacon](#)). Cependant, dans le contexte actuel de Beacon v2, les deux éléments sont considérés interdépendants et seront probablement mis à jour en même temps pour les principales versions. Voir : <https://docs.genomebeacons.org/what-is-beacon-v2/?h=beacon>

Figure 3 : Illustration du Framework et du Modèle Beacon



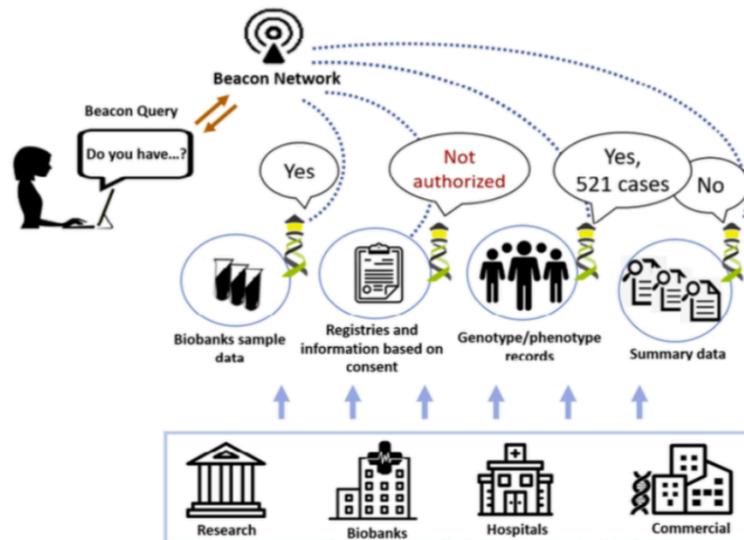
Source : [Rambla et al. \(2022\)](#).²

- Les **instances Beacon** peuvent être combinées de différentes manières :
 - Un **réseau Beacon** est une combinaison de plusieurs instances Beacon individuelles dans une même interface, potentiellement issues de plusieurs institutions ou producteurs de données différents⁹. Dans un réseau Beacon, l'accès est uniformisé à travers une interface personnalisée (voir Figure 4), et éventuellement grâce à un alignement des paramètres des instances (ex : filtres harmonisés).
 - Des **instances Beacon** peuvent également être combinées à travers des **Beacon Aggregator** : il s'agit d'interfaces d'accès uniques dans lesquelles les instances Beacons individuelles ne sont cependant ni nécessairement harmonisées, ni nécessairement averties de leur intégration dans l'agrégateur¹⁰.

9) Voir : [Beacon Networks and Aggregators - Beacon v2 Documentation](#)

10) Dans le cadre des développements liés à la médecine personnalisée, l'utilisation des données génomiques est envisagée pour prédire l'évolution de la santé des individus dans un objectif de prévention et de prise en charge individualisée. Voir l'article de Prins BP, Leitsalu L, Pärna K, Fischer K, Metspalu A, Haller T, Snieder H. « Advances in Genomic Discovery and Implications for Personalized Prevention and Medicine: Estonia as Example ». J Pers Med. 2021 Apr; [Advances in Genomic Discovery and Implications for Personalized Prevention and Medicine: Estonia as Example - PMC](#)

Figure 4 : Exemple d'un réseau Beacon



Source : [Beacon Networks and Aggregators - Beacon v2 Documentation](#)

- **Organisme en charge** : Global Alliance for Genomics and Health (GA4GH).

APPLICATION

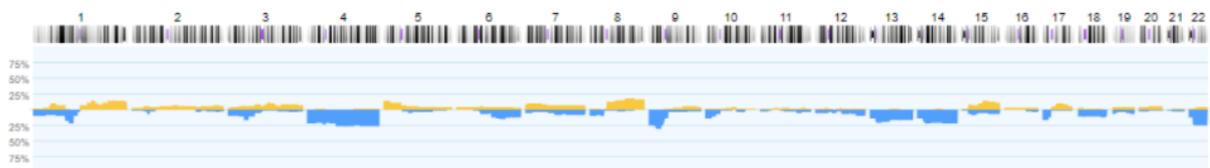
- **Domaine d'application en santé** : recherche **génomique** et **clinique** ³.
- **Principaux cas d'usage** : **identification de gènes** associés à des maladies, en particulier dans le cas du cancer (gene or genetic discovery)¹⁰, **identification de cohortes** de patients, **découverte de données génomiques** (data discovery) pour la recherche biomédicale.
- **Illustration concrète, exemple d'utilisation sur un cas simple** :
 - La base de données **Progenetix Cancer Genomics Information Resource**, créée par le Computational Oncogenomics Group de l'Université de Zürich et du Swiss Institute of Bioinformatics (SIB), contient les **profils génomiques** de plus de **140 000 individus** atteints de **cancer** ¹¹.
 - Cette base permet de disposer d'un aperçu des **mutations et variations génétiques**, pour tous les types de tumeurs. Depuis 2016, Progenetix implémente une **version expérimentale de Beacon** dans laquelle de nouveaux concepts sont développés pour étendre le protocole Beacon au-delà de la spécification de la version 2 (options de requêtage CNV, transmission de données) ¹². Ainsi, il est possible de **requêter la base de données Progenetix**, en effectuant un filtre sur le sexe génotypique, le type de cancer, le stade du cancer, le type de mutation, le chromosome, le locus du gène sur un chromosome, etc.

11) [Other implementations - Beacon v2 Documentation](#) et [Progenetix](#)

12) [Beacon +](#)

- Les **technologies** utilisées sont les suivantes :
 - MongoDB backend.
 - Python-based full stack API / middleware (bycon)¹³.
 - React based front-end (progenetix-web)¹⁴.
- L'interface Beacon+¹⁵ de Progenetix permet par exemple d'afficher la fréquence de l'augmentation et de la diminution du nombre de copies dans 22 autosomes (voir Figure 5).

Figure 5 : Illustration des résultats obtenus à l'aide de la plateforme Beacon Progenetix



Source : [Progenetix](#), Kon et al. (2013)¹⁶

DONNÉES

- **Typologie de données concernées** : identifiants de l'échantillon biologique, de l'analyse, de l'individu concerné, méthodologie utilisée, plateforme utilisée, date de l'analyse, liste de maladies de l'individu, mode de vie de l'individu (habitudes, comportements), etc. (voir l'élément « Complexité du modèle » ci-dessous pour la description précise des tables du Modèle et des données associées).
- **Type de granularité** : les données peuvent concerner soit des individus appartenant à une population d'étude (dans la table Cohorts, à laquelle sont liées les tables Individuals, Biosamples, Runs, Analyses, GenomicVariations, voir l'élément « Complexité du modèle » ci-dessous), soit des variants regroupés dans une base de données (dans la table Datasets, à laquelle est liée la table GenomicVariations).

UTILISATION DANS PLUSIEURS LANGUES

La spécification Beacon et la documentation associée sont en **anglais**.

13) [GitHub - progenetix/bycon: Bycon - A Python Based Beacon API \(beacon-project.io\) implementation leveraging the Progenetix \(progenetix.org\) data model](#)

14) [Web frontend for the Progenetix cancer genome resource](#)

15) [Beacon +](#)

16) Kon, A., Shih, LY., Minamino, M. et al. « Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms ». Nat Genet 45, 1232–1237 (2013). [Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms | Nature Genetics](#)

DISPONIBILITÉ DE LA DOCUMENTATION D'IMPLÉMENTATION

- GitHub unifié qui contient¹⁷ :
 - GitHub du Framework¹⁸ Beacon.
 - GitHub du Modèle¹⁹ Beacon.
- Documentation Beacon v2²⁰.
- Présentation du projet Beacon³.

NIVEAU DE GÉNÉRALISATION (FACILITÉ DE REMPLISSAGE DES CHAMPS DU STANDARD)

Note : 0,7 / 1. Cette note combine plusieurs sous-critères. Elle s'explique par :

- a)** Le fait que les terminologies ne sont pas imposées (voir l'élément « Flexibilité dans les choix des terminologies » en partie 3. Technique).
- b)** La flexibilité du standard (voir l'élément « Flexibilité du standard, personnalisation » en partie 3. Technique).
- c)** L'absence de contraintes d'implémentation (voir l'élément « Contraintes d'implémentation » en partie 3. Technique).
- d)** Une couverture relativement faible des cas d'usages (voir l'élément « Principaux cas d'usages » en partie 1. Général).
- e)** Une couverture relativement faible des domaines d'application (voir l'élément « Domaine d'application en santé » en partie 1. Général).

En raison de la flexibilité du standard et des terminologies, la spécification **peut être facilement généralisée et utilisée dans d'autres domaines d'application** (voir les éléments « Flexibilité du standard, personnalisation » et « Flexibilité dans les choix des terminologies » en partie 3. Technique).

17) [Unified repository for the GA4GH Beacon v2 API standard](https://github.com/ga4gh-beacon/beacon-v2/tree/main/framework/src)

18) <https://github.com/ga4gh-beacon/beacon-v2/tree/main/framework/src>

19) <https://github.com/ga4gh-beacon/beacon-v2/tree/main/models/src>

20) [Beacon Documentation](https://github.com/ga4gh-beacon/beacon-v2/tree/main/docs) ou ici : <https://github.com/ga4gh-beacon/beacon-v2/tree/main/docs>

2. Gouvernance



MODALITÉS D'ACCÈS ET DISTRIBUTION DES SOLUTIONS BASÉES SUR CE STANDARD

- Le **standard Beacon v2** est publié sous la licence **Creative Commons Zero v1.0 Universal**²¹.
- Le **logiciel B2RI** est publié sous la **licence GNU General Public License v3.0 et Apache License v2.0**²².

GOVERNANCE DES ÉVOLUTIONS DU STANDARD

- Le standard est développé par le groupe de travail **Discovery de la GA4GH**²³. Il est possible de rejoindre ce groupe de travail à **différents niveaux d'intervention** : pour apporter son feedback sur les documents, participer aux rendez-vous réguliers, rédiger les documents et coder le standard²⁴.
 - Au sein de la GA4GH, le coordinateur du projet Beacon est Amreen Mohamed, et les responsables du produit sont Anthony Brookes, Jordi Rambla et Michael Baudis²⁴.
 - Des contributions additionnelles ont également été reçues de la part d'organisations extérieures, de développeurs individuels et d'utilisateurs²⁵. L'ensemble des contributions et des participants au projet Beacon ainsi que leurs contacts sont référencés²⁶.
- **ELIXIR**, l'infrastructure européenne de bio-informatique, et **GA4GH** travaillent en commun au développement du projet Beacon :
 - Une première collaboration a donné lieu au développement de **six instances Beacon en Europe** (Suède, Finlande, France, Suisse, Belgique et European Genome-phenome Archive ou EGA, un projet entre EMBL-EBI et le Center for Genomic Regulation à Barcelone) à travers les nœuds ELIXIR (ELIXIR Nodes : il s'agit du rassemblement d'instituts de recherche au sein d'un pays européen, ils constituent le réseau ELIXIR)²⁷.

21) <https://github.com/ga4gh-beacon/beacon-v2/blob/main/LICENSE>

22) Article de Rueda et al., « Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data », Bioinformatics, Volume 38, Issue 19, October 2022, Pages 4656–4657, [Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data | Bioinformatics | Oxford Academic](#) et [B2RI Documentation](#)

23) [Beacon – GA4GH](#). Il existe plusieurs groupes de travail permettant de contribuer au développement des produits GA4GH. Pour les rejoindre, compléter le formulaire : [Join our Work Streams and communities – GA4GH](#)

24) [Join our Work Streams and communities – GA4GH](#)

25) [How to Contribute to Beacon Development - Beacon v2 Documentation](#)

26) <https://beacon-project.io/people/?h=contact>

27) [The ELIXIR Beacon - 2017](#)



- En 2017, ils ont annoncé une **extension de leurs objectifs de collaboration avec les ambitions suivantes**²⁸ : (1) établir un réseau de Beacon ELIXIR²⁹ (2) développer de nouvelles fonctionnalités (3) renforcer la sécurité (4) nouer des partenariats stratégiques avec des propriétaires de données nationales.
- Une demande d'adhésion au réseau Beacon ELIXIR peut être faite en remplissant un formulaire³⁰.

MATURITÉ DU STANDARD

- **Fréquence de mise à jour : la version 1** de la spécification Beacon a été publiée **en août 2021**³¹. **La version 2** a été publiée **en février 2022**³² et approuvée en avril 2022 comme un **standard GA4GH officiel**³³.

28) Voir : [ELIXIR and GA4GH Beacon Team Up to Advance Genomic Data Sharing](#).

29) Voir le réseau de Beacons ELIXIR à date : [ELIXIR Beacon Network](#)

30) Voir le formulaire : <https://beacon-network.elixir-europe.org/join>

31) Voir le GitHub : [GitHub - ga4gh-beacon/Model-BEACON-v1](#)

32) Il s'agit de la date de la première publication (commit) sur le GitHub de Beacon v2 : https://github.com/ga4gh-beacon/beacon-v2/commits/main?after=e7d79620cccc3c93e9a559d9add3955491a6614c+384&branch=main&qualified_name=refs%2Fheads%2Fmain

33) Voir : [GitHub - ga4gh-beacon/Model-BEACON-v1](#)

3. Technique



INTEROPÉRABILITÉ

- Le modèle Beacon est interopérable avec les modèles OMOP, FHIR et OpenEHR³⁴ :
 - Le modèle Beacon est compatible avec les standards sémantiques utilisés pour décrire les concepts cliniques.
 - Ainsi qu'avec les standards syntaxiques disponibles pour l'harmonisation du stockage et du partage des données cliniques : le modèle de données "individuals" Beacon est mappé avec les standards OMOP, FHIR et OpenEHR, et son API permet de requêter des données structurées sous ces trois standards.

FLEXIBILITÉ DANS LES CHOIX DES TERMINOLOGIES

- La spécification Beacon permet la **flexibilité dans le choix des terminologies en proposant une approche d'harmonisation** au moment de la requête² :
 - Le Framework Beacon v2 suggère l'utilisation d'un terme d'ontologie au moment de la requête (ex : pour représenter le concept « homme », les termes utilisés sont identifiants dans l'ontologie PATO, PATO:0000384).
 - Les instances Beacon traduisent et alignent les termes d'ontologie fournis dans la requête vers leurs représentations internes correspondantes, assurant ainsi une harmonisation lors de la requête et de la réponse. Ces représentations internes peuvent être différentes entre les instances (ex : pour représenter le concept « homme », une instance pourrait utiliser le vocabulaire « M », tandis que l'autre utiliserait le concept « male »).
- Beacon n'impose **pas de contraintes sur les terminologies** qui peuvent être utilisées : **la spécification est agnostique** vis-à-vis des normes sémantiques utilisées par les sources de données. Il est compatible avec de nombreuses ontologies biomédicales standards telles que SNOMED CT pour les diagnostics, HPO pour les anomalies phénotypiques et LOINC pour les résultats de laboratoire².
- Beacon utilise la spécification Phenopackets v2³⁵ pour décrire et représenter les ontologies².

34) [Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond - Rambla - 2022 - Human Mutation - Wiley Online Library](#)

35) Il s'agit d'un standard approuvé par GA4GH pour partager des informations sur les maladies et les phénotypes, et a été adopté par la communauté de recherche sur les maladies rares pour caractériser et représenter les manifestations des maladies. Voir l'article de Rubinstein, Y. R., et al. (2020). « The case for open science : Rare diseases. »

JAMIA Open, 3(3), 472–486. [The case for open science: rare diseases](#)

FLEXIBILITÉ DU STANDARD, PERSONNALISATION

La spécification Beacon v2 **utilise un Modèle de données extensible** ainsi qu'un Framework flexible² :

- La séparation du Framework et du Modèle Beacon permet d'adapter le concept Beacon à d'autres disciplines sans s'éloigner de la spécification initiale : **Beacon v2 est agnostique au domaine d'application**. En particulier :
 - Le Framework est agnostique au domaine d'application car il contient uniquement la structure des requêtes API et des réponses.
 - La spécification Beacon v2 contient un Modèle par défaut qui propose une structure recommandée pour les données génomiques et cliniques, mais qui n'est cependant pas imposée.
- Une instance Beacon peut être implémentée au-dessus d'une architecture existante, comme c'est le cas dans plusieurs exemples d'implémentations (Cafe Variome, OpenCGA, Progenetix, RD-Connect GPAP, voir l'élément « Adoption du standard » en partie 4. Valorisation).

COMPLEXITÉ DU MODÈLE

Ce Modèle par défaut comporte les 7 entités suivantes qui sont liées par 12 relations et contiennent les termes suivants^{36,37} :

- Les entités **Datasets** et **Cohorts** (les collections) : contiennent un regroupement de variants d'une même base de données (dans **Datasets**³⁸ : date de création, conditions d'utilisation, description, identifiant, nom, version, etc.) ou d'individus appartenant à la même population d'étude (dans **Cohorts**³⁹ : design de la cohorte, taille, type, identifiant, critères d'inclusion, nom, etc.).
- L'entité **Genomic variation**⁴⁰ : contient les informations sur les altérations génomiques uniques (ex : position dans le génome, altérations des séquences, type, fréquence dans la population, etc.).
- L'entité **Individuals**⁴¹ : contient les détails phénotypiques et cliniques des patients ou des individus sains constituant les groupes de contrôle (ex : maladies, origine ethnique, mode de vie, origine géographique, interventions cliniques, traitements, caractéristiques phénotypiques, sexe, etc.).
- L'entité **Biosamples**⁴² : contient les détails des échantillons prélevés sur les individus (ex : date de collecte, âge de l'individu à la collecte, diagnostic à partir des marqueurs, diagnostic histologique, progression de la tumeur, etc.).

36) [beacon-v2-Models](#)

37) [Beacon terms - Beacon v2 Documentation](#) et le [GitHub des éléments communs aux tables](#)

38) [datasets defaultSchema - Beacon v2 Documentation](#) et le [GitHub associé](#)

39) [cohorts defaultSchema - Beacon v2 Documentation](#) et le [GitHub associé](#)

40) [genomicVariations defaultSchema - Beacon v2 Documentation](#) et le [GitHub associé](#)

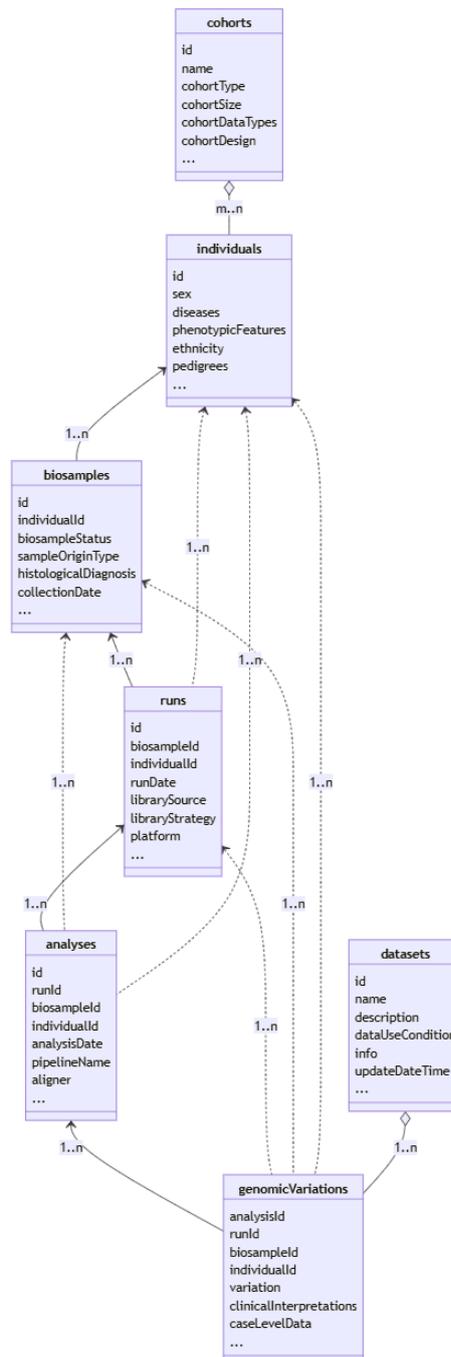
41) [individuals defaultSchema - Beacon v2 Documentation](#) et le [GitHub associé](#)

42) Voir le contenu de l'entité Biosamples : [biosamples defaultSchema - Beacon v2 Documentation](#) et le [GitHub associé](#)

- Les entités **Analyses** et **Runs** : contiennent les détails sur :
 - Les procédures utilisées pour séquencer un échantillon biologique (**Runs**⁴³ : méthodologie et plateforme technologique utilisées, etc.).
 - Les procédures bio-informatiques utilisées pour identifier les variants (**Analyses**⁴⁴ : date de l'analyse, pipeline utilisé, etc.).

Les liens entre ces entités sont illustrés dans le schéma ci-dessous (voir Figure 6 ci-dessous).

Figure 6 : Modèle Beacon



Source : [Beacon v2 documentation](#)³⁷

43) Voir le contenu de l'entité Runs : [Beacon v2 Documentation](#) et le [GitHub associé](#)

44) Voir le contenu de l'entité Analyses : [Beacon v2 Documentation](#) et le [GitHub associé](#)

TECHNOLOGIE DE STOCKAGE ET TRAITEMENT DE DONNÉES ET NIVEAU D'ADOPTION DE LA TECHNOLOGIE

Le standard n'impose pas de technologie de stockage ou de traitement des données.

NEUTRALITÉ TECHNOLOGIQUE : OUI

Le Modèle et le Framework sont décrits aux formats YAML et JSON mais les modalités techniques de mise en place ne sont pas imposées³⁷.

COMPÉTENCES TECHNIQUES ET MÉTIER NÉCESSAIRES POUR UTILISER LE STANDARD

- Pour mettre en place la spécification à partir du logiciel B2RI²³ (Beacon v2 Reference Implementation), les compétences suivantes sont nécessaires :
 - **Compétences techniques :**
 - Système d'exploitation Linux.
 - Langages de programmation Python, Perl.
 - Base de données NoSQL MongoDB.
 - Services Web et API.
 - Docker pour la version containerisée.
 - Ces compétences techniques correspondent à des profils Software Engineer, Data Engineer, ingénieur des systèmes d'information.
 - Pour aligner l'API Beacon avec le contenu de la base de données, des **compétences métier** sur la génomique sont nécessaires.
- Pour **utiliser une interface de recherche basée sur la spécification Beacon**, il n'y a **pas de compétences requises**.

ÉCHANGE SÉCURISÉ DE DONNÉES

- Beacon autorise trois niveaux de sécurité pour l'accès aux données⁴⁵ :
 - Niveau 1 : **accès anonyme**. Le Beacon répond aux requêtes quelle qu'en soit la source.
 - Niveau 2 : **accès d'un tiers enregistré**. L'utilisateur doit s'enregistrer, par exemple en utilisant une identité ELIXIR.
 - Niveau 3 : **accès contrôlé**. L'utilisateur doit avoir demandé l'autorisation et l'avoir reçue avant de faire sa requête.

45) [Frequently Asked Questions - Beacon v2 Project Website](#)



- Au sein d'un Beacon, il peut y avoir plusieurs niveaux d'autorisation selon la base de données concernée, par exemple pour respecter les conditions de collecte des données. Le niveau d'accès peut également dépendre de la requête et de l'utilisateur.
 - Les attributs de sécurité font partie du Framework Beacon : le fichier `beaconConfiguration.json`⁴⁶ définit le schéma du fichier JSON qui inclut les aspects fondamentaux de la configuration. La troisième section (`securityAttributes`) définit la sécurité.
 - Un document décrivant les bonnes pratiques de sécurité a également été rédigé⁴⁷.

46) Voir : [beacon-v2/framework/json/configuration/beaconConfigurationSchema.json at main](#)

47) Voir : [ELIXIR Beacon 2019-21 Deliverable D3.3](#)

4. Valorisation



ACCESSIBILITÉ À DES RESSOURCES DE FORMATION

- Webinaires et vidéos disponibles sur la chaîne YouTube de GA4GH⁴⁸.

DISPONIBILITÉ DE LA DOCUMENTATION SCIENTIFIQUE DÉMONSTRANT L'INTÉRÊT

- En juin 2023, une **dizaine d'articles**⁴⁹ traitant de Beacon sont disponibles sur PubMed (voir la requête et un extrait de la liste des publications en « Annexe n°1 »). Ces articles abordent principalement deux thématiques :
 - Les **enjeux de sécurisation et de confidentialité des données génomiques**⁵⁰.
 - Le **fonctionnement de Beacon, son rôle et son intérêt dans la collecte, le partage et l'utilisation des données génomiques**. Selon Fiume et al. (2019)⁵¹, le projet Beacon montre, dès 2019, la volonté de divers organismes de coopérer, de définir des normes, de s'engager dans l'échange et dans l'analyse des données génomiques.
 - En effet, Beacon a été implémenté par des petites comme par des grandes organisations, tout comme par des chercheurs à titre individuel. Cela a permis de collecter des séquençages à grande échelle.
 - Le partage de données génomiques par l'intermédiaire de Beacon diffère du partage de données génomiques par des formats plus classiques (ex : VCF, GFF), en ce que Beacon s'occupe de l'agrégation des données et du « brouillage » des métadonnées pour que le lien entre séquences génomiques et des individus spécifiques ne puisse pas être fait.
 - Le partage des données génomiques sera permis grâce aux nœuds ELIXIR (voir l'élément « Processus de prise de décision sur le standard » en partie 2. Gouvernance). Avec une adoption croissante de Beacon, un vaste ensemble de données pourra être constitué à l'échelle mondiale. L'analyse de ces données génomiques peut conduire à de nouvelles découvertes et applications médicales.

48) Voir les vidéos : [WEBINAR - Using the GA4GH toolkit: Beacon API V1.0.0 for genomic data discovery](https://www.youtube.com/watch?v=WWMHgYrPWHg&pp=ygUMROEOR0ggQkVBQ09Q), [https://www.youtube.com/watch?v=WWMHgYrPWHg&pp=ygUMROEOR0ggQkVBQ09Q](https://www.youtube.com/watch?v=W-sSLgXGP9s&pp=ygUMROEOR0ggQkVBQ09Q) ; <https://www.youtube.com/watch?v=W-sSLgXGP9s&pp=ygUMROEOR0ggQkVBQ09Q> , [GA4GH Connect: Beacon API](https://www.youtube.com/results?search_query=GA4GH+BEACON) et autres (https://www.youtube.com/results?search_query=GA4GH+BEACON)

49) L'estimation du nombre d'articles traitant de Beacon sur PubMed a été faite en saisissant les termes « Beacon » et « GA4GH » dans la barre de recherche du site PubMed ; voir : <https://pubmed.ncbi.nlm.nih.gov/?term=Beacon+GA4GH>. Voir en Annexe n°1 pour les détails de la requête PubMed et des résultats obtenus.

50) Voir par exemple l'article de Raisaro JL, Tramèr F, Ji Z, Bu D, Zhao Y, Carey K, Lloyd D, Sofia H, Baker D, Flicek P, Shringarpure S, Bustamante C, Wang S, Jiang X, Ohno-Machado L, Tang H, Wang X, Hubaux JP. « Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks ». J Am Med Inform Assoc. 2017 : [Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks](#)

51) Voir l'article de Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, Dyke SOM, Brookes AJ, Carey K, Lloyd D, Goodhand P, Haeussler M, Baudis M, Stockinger H, Dolman L, Lappalainen I, Törnroos J, Linden M, Spalding JD, Ur-Rehman S, Page A, Flicek P, Sherry S, Haussler D, Varma S, Saunders G, Scollen S. « Federated discovery and sharing of genomic data using Beacons ». Nat Biotechnol. 2019 : [Federated discovery and sharing of genomic data using Beacons - PMC](#)



ADOPTION DU STANDARD

En juin 2023, il n'y a pas d'adoption officielle de la spécification Beacon.

- Utilisation sur le marché :

En 2019, **42 organisations internationales** avaient **implémenté l'API Beacon** sur plus de **200 bases de données** contenant les données relatives à plus de 100 000 individus anonymisés (voir Figure 7)⁴.

- Le **Beacon Network**⁵², réseau de Beacon publics à travers le monde, **a été utilisé 1,5 million de fois** pour des requêtes et constitue à présent une ressource internationale d'envergure pour la découverte de données facilitée.
- Les archives génomiques européennes et américaines ont mis en place des instances Beacon : la base de données américaine des génotypes et phénotypes⁵³ (NIH dbGaP), la European Genome-phenome Archive⁵⁴ (EGA, au sein du réseau ELIXIR, voir ci-dessous) et la European Variation Archive⁵⁵ (EVA).
- L'infrastructure européenne **ELIXIR** de bio-informatique a permis la diffusion de l'utilisation de la spécification Beacon en Europe et la constitution d'un **réseau ELIXIR d'utilisateurs de l'API Beacon**⁵⁶ qui comprend notamment les institutions suivantes (voir l'élément « Processus de prise de décision sur le standard » en partie 2. Gouvernance) :
 - L'Université de Zurich et l'Institut Suisse de Bio-informatique (SIB : Swiss Institute of Bioinformatics) : Baudisgroup (Progrenetix Cancer Genomics Beacon+, voir l'élément « Illustration concrète, exemple d'utilisation sur un cas simple » en partie 1. Général).
 - L'Institut de génomique à l'Université de Tartu en Estonie.
 - L'Infrastructure nationale de bio-informatique en Suède (NBIS : National Bioinformatics Infrastructure Sweden).
 - Une plateforme de recherche biomédicale à Barcelone en Espagne (DisGeNET).
 - L'Institut européen de bio-informatique (EMBL : European Molecular Biology Laboratory).
 - L'Université du Luxembourg (LCSB : Luxembourg Centre for Systems Biomedicine).
 - Le Centre de technologie de l'information pour la science de Finlande (CSC – IT Center for Science).
 - L'institut national de la santé et du bien-être en Finlande.

52) [Beacon Network](#)

53) [Home - dbGaP - NCBI](#)

54) <https://ega-archive.org/beacon/#/>

55) [EBI GA4GH Beacon](#)

56) [ELIXIR Beacon Network](#)



- En dehors du réseau ELIXIR, l'API Beacon a également été implémentée dans un large éventail de cas d'utilisation :
 - **The 1000 Genomes** : données de séquençage de population à grande échelle.
 - **Poly-Phen** : prédictions in silico pour le diagnostic clinique.
 - **Base de données sur les mutations génétiques humaines (HGMD : Human Gene Mutation Database)** : données provenant de bases de données créées par des experts ou issues de la compilation de nombreuses bases de données.
 - **ClinVar** : agrégation d'informations sur les variations génomiques et leur implication en termes de santé humaine.
 - **L'International Cancer Genome Consortium (ICGC)** : variants somatiques dans plus de 60 sous-types de cancer.
 - **BRCA Exchange** : groupe de travail pour créer des classifications consensuelles pour les variants de BRCA1 et BRCA2.
- Des organisations tierces ont également permis la connexion de leurs propres bases de données à des données de variations génétiques à travers l'API Beacon : Cafe Variome, DNASTack, Global Gene Corp, Genecloud, Google Cloud.

FOURNISSEURS DE SERVICE AYANT L'EXPERTISE EN FRANCE

- ELIXIR France⁵⁷, nœud français d'ELIXIR en relation avec le hub européen d'ELIXIR à Hinxton.

57) Voir : [ELIXIR France](#)

5. Utilisation



SIMPLICITÉ D'USAGE

- **Note : 0,9 / 1.** Cette note combine plusieurs sous-critères. Elle s'explique par :

- a)** L'accès à des ressources officielles de formation (voir l'élément « Accessibilité à des ressources de formation » en partie 4. Valorisation).
- b)** La lisibilité du schéma par un humain (voir l'élément « Lisible par un humain » en partie 5. Utilisation).
- c)** Le nombre relativement faible de profils requis pour l'implémentation et l'usage (voir l'élément « Compétences techniques et métier nécessaires pour utiliser le standard » en partie 3. Technique).

- La spécification Beacon est **simple à utiliser** car, une fois mise en place, elle consiste en une interface web sur laquelle l'utilisateur peut directement faire ses requêtes : l'exemple de l'interface Beacon+ Progenetix illustre cette simplicité d'usage.
- De plus, la flexibilité du standard et la possibilité de **l'implémenter au-dessus d'une architecture existante facilite sa mise en place** (voir l'élément « Flexibilité du standard, personnalisation » en partie 3. Technique).

EXISTENCE D'UNE COMMUNAUTÉ EN LIGNE ET DEGRÉ D'ACTIVITÉ

- GA4GH Beacon-network Google Group⁵⁸.

OUTILS DE MAPPING

- En juin 2023, nous n'avons pas identifié d'outils de mapping.

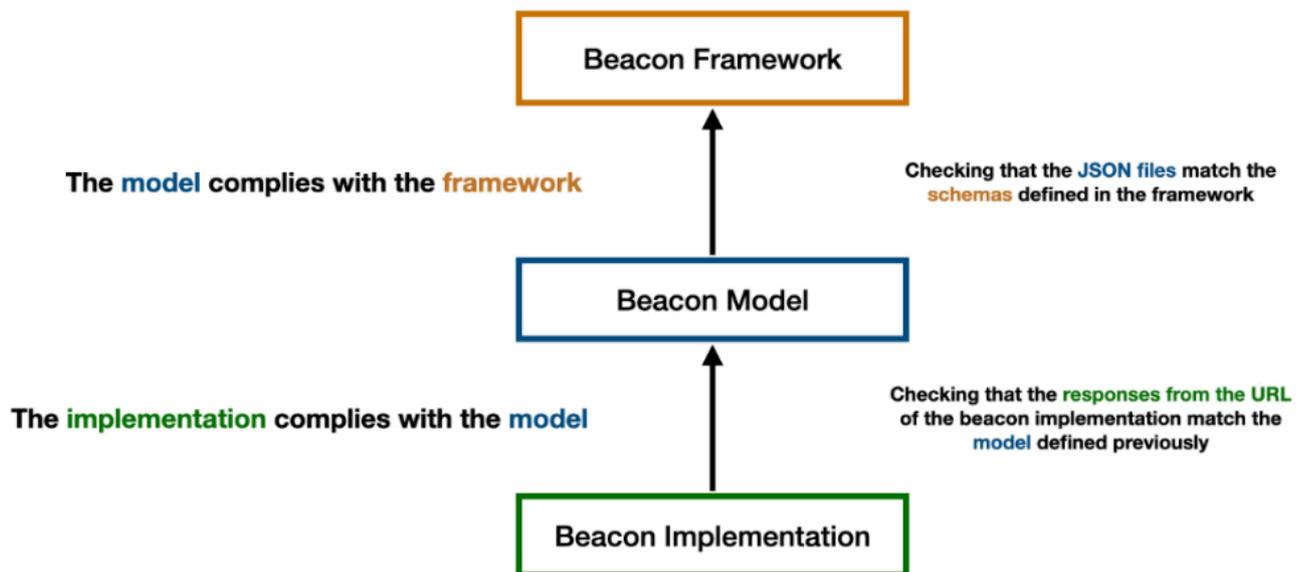
OUTILS COMPATIBLES

- **Beacon verifier** : outil de vérification de la conformité de l'implémentation de Beacon au Modèle et au Framework (voir Figure 6)⁵⁹.

58) [GA4GH Beacon Network - Google Groups](#) . En juin 2023, la dernière activité date du 26 octobre 2022. L'activité du réseau est relativement faible avec environ 2 discussions par an.

59) [GitHub - ga4gh-beacon/beacon-verifier: Tool to verify that a Beacon implementation follows the specification](#)

Figure 7 : Fonctionnement de l'outil Beacon (verifier)



Source : [GitHub du Beacon verifier](#)⁶⁰

- Outils d'aide à l'implémentation de Beacon contenus dans le logiciel B2RI (Beacon v2 Reference Implementation)⁶⁰ :
 - **BFF Validator** (Beacon Friendly Format Validator) : outil qui vérifie l'alignement entre la structure des données source et le Modèle JSON et, en cas de validation, qui crée un code JSON en sortie.
 - **BFF Queue** : outil qui permet de transformer les données génomiques VCF en entité GenomicVariations et de créer un fichier JSON en sortie, de paralléliser le traitement des fichiers et de charger les données dans une instance MongoDB.
 - **BFF API** : API de démonstration qui permet de réaliser les requêtes élémentaires auprès d'une instance MongoDB.

60) Repository GitHub : <https://github.com/EGA-archive/beacon2-ri-tools/tree/main/utlis>

DÉCRIRE LES ÉTAPES NÉCESSAIRES POUR LA STANDARDISATION

Pour implémenter une instance Beacon locale, deux éléments sont nécessaires :

- a) Une base de données contenant les données biologiques.
- b) Une API REST qui permet d'envoyer des requêtes et de recevoir des réponses de manière standardisée. A cette fin, B2RI²³ (Beacon V2 Reference Implementation) fournit ces deux éléments essentiels ainsi qu'un ensemble d'outils permettant de "beaconiser" les données. En particulier, B2RI contient 4 composantes :
 - 1) Un ensemble d'outils d'extraction, de transformation et de chargement des métadonnées (méthodologie de séquençage, outils de bio-informatique, etc.), des données phénotypiques et des variants génomiques dans une base de données.
 - 2) La base de données (une instance MongoDB).
 - 3) L'API REST Beacon V2.
 - 4) Un exemple de base de données (données synthétiques).

Pour utiliser B2RI, il faut suivre les étapes suivantes :

- 1) Installer le logiciel, soit depuis une image Docker ou depuis GitHub (voir les recommandations pour l'installation de B2RI dans le Supplementary Text ST7). Le logiciel doit être installé en local sur un **serveur Linux**.
- 2) Transformer les métadonnées et les données phénotypiques vers la structure hiérarchique du Modèle Beacon v2 (voir l'élément « Complexité du modèle » pour la description des 7 entités) au format JSON. Pour « beaconiser » les données, B2RI contient un outil de validation de leur structure (BFF Validator).
- 3) Charger les fichiers JSON issus du BFF Validator dans une base de données.
- 4) Transformer les données des variations génomiques avec l'outil B2RI dédié (BFF Queue).
- 5) Charger les 7 fichiers JSON (ils constituent le BFF : Beacon Friendly Format) dans une instance MongoDB à l'aide de l'outil BFF Queue. Lorsqu'elles sont chargées dans la base de données, les entités sont appelées des « collections » MongoDB. À ce stade, les données peuvent être requêtées et n'ont pas besoin d'être à nouveau mappées à une API. Si cette étape de chargement dans une instance MongoDB n'est pas réalisée, il faut cependant réaliser un mapping entre la base de données et l'API Beacon.



- 6)** Configurer l'API Beacon (en particulier en termes de sécurité et de granularité des réponses).
- 7)** Déployer l'API Beacon. Les requêtes sont envoyées en utilisant les méthodes GET ou POST HTTP à des endpoints de l'API Beacon v2. Les requêtes utilisent des request parameters qui permettent de mapper le vocabulaire de l'API aux collections MongoDB. Elles peuvent ensuite être affinées en utilisant des filtering terms (Bio-ontology, Custom, Numeric, Alphanumeric). Une API Beacon de démonstration est disponible (BFF API).

EXISTENCE D'EXTENSIONS CERTIFIÉES

En juin 2023, il n'existe pas d'extensions certifiées.

BIBLIOTHÈQUE DE REQUÊTES TYPES

En juin 2023, il n'existe pas de bibliothèques de requêtes types.

LISIBLE PAR UN HUMAIN : OUI

Les données sont transférées en format JSON qui est lisible par humain.

6. Glossaire



1. **API** : Application Programming Interface
2. **B2RI** : Beacon v2 Reference Implementation
3. **BFF** : Beacon Friendly Format
4. **BRCA 1 / 2** : BReast CAncer 1 / 2
5. **CNA** : Copy Number Abbreviation
6. **CNV** : Copy Number Variation
7. **CSC-IT** : Center for Science – IT (IT : Information Technology)
8. **EBI** : European Bioinformatics Institute
9. **EGA** : European Genome-phenome Archive
10. **EVA** : European Variation Archive
11. **EHR** : Electronic Health Record
12. **EMBL** : European Molecular Biology Laboratory
13. **ENIGMA** : Enhancing Neuro Imaging Genetics Through Meta Analysis
14. **FHIR** : Fast Healthcare Interoperability Resources
15. **GA4GH** : Global Alliance for Genomics and Health
16. **GFF** : General Feature Format
17. **GNU** : GNU is Not Unix
18. **GPAP** : Genome-Phenome Analysis Platform
19. **HGMD** : Human Gene Mutation Database
20. **HTTP** : Hypertext Transfer Protocol
21. **ICGC** : International Cancer Genome Consortium
22. **JSON** : JavaScript Object Notation
23. **LCSB** : Luxembourg Centre for Systems Biomedicine
24. **NBIS** : National Bioinformatics Infrastructure Sweden
25. **NIH dbGaP** : National Institutes of Health database of Genotypes and Phenotypes
26. **OMOP-CDM** : Observational Medical Outcomes Partnership Common Data Model
27. **REST** : REpresentational State of Transfer
28. **SIB** : Swiss Institute of Bioinformatics
29. **SQL** : Structured Query Language
30. **VCF** : Variant Call Format
31. **YAML** : Yet Another Markup Language