HEALTH DATA HUB

OSIRIS

GrOupe inter-SIRIC sur le paRtage et l'Intégration des donnéeS clinico-biologiques en cancérologie

CIM-11
Classification Internationale des Maladies et des problèmes de santé connexes

Ce document a été réalisé en partenariat avec VELTYS.

Description et type de standard : OSIRIS (GrOupe inter-SIRIC sur le paRtage et l'Intégration des donnéeS clinico-biologiques en cancérologie) est un **schéma de données** théorique et un **ensemble minimal de données** qui permet de représenter les **données cliniques et génomiques de patients atteints de cancer** à travers une succession d'**événements** (antécédents, diagnostics, traitements, réactions, analyses, etc.). Il s'agit d'un standard français créé par l'Institut National du Cancer (INCa)[1].

Domaine d'application : Oncologie.

Maturité/utilisation : OSIRIS est un standard récent, dont la première version (**v1**) a été publiée en 2018 et la version actuelle (**v1.1.05**) est sortie en février 2019. À ce jour, il est utilisé par les institutions membres de l'initiative OSIRIS (Institut Curie, Institut Bergonié, Centre Léon Bérard, Institut du Cancer de Montpellier, Institut Paoli-Calmettes, Institut Gustave Roussy, CHU de Bordeaux, Hôpital Européen Georges Pompidou, Hôpital Saint Louis, Unicancer, voir l'élément « Adoption du standard » en partie 4. Valorisation).

1) Voir : OSIRIS - Une approche « open science » et conceptuelle pour l'analyse de données interopérables en oncologie - Recherche translationnelle

1. Général

PRÉSENTATION

- **Pays d'origine** : France.
- **Consortium d'origine** : développée par **l'Institut National du Cancer** (INCa)[2].
- **Type de standard** : schéma de données et ensemble minimum de données.
- **Description** :
  - OSIRIS est un **schéma de données** et un **ensemble minimal de données cliniques et omiques** lancé en 2015 par l'Institut National du Cancer (INCa) en France[3]. Il a pour objectif de permettre aux chercheurs et aux médecins de disposer dans le domaine de l'oncologie (1) de **données homogènes** (2) qui permettent de capter l'évolution de la maladie et en particulier la résistance aux interventions thérapeutiques et la toxicité. En particulier, il a été développé pour pouvoir analyser de manière conjointe les données issues de plusieurs **essais cliniques** menés dans les Sites de Recherche Intégrée sur le Cancer (SIRICs) en France.
  - L'initiative OSIRIS propose de traiter le problème de l'hétérogénéité des données en prenant les 4 engagements suivants :

    1) Définir un ensemble minimal de données aussi restreint que possible.

    2) Atteindre un consensus national entre tous les acteurs intervenants dans la recherche sur le cancer.

    3) Utiliser des terminologies internationales aussi établies que possible.

    4) Définir des règles d'implémentation qui garantissent la cohérence de l'ensemble minimal de données dans les différentes institutions.
  - Il existe plusieurs versions d'OSIRIS en cours d'élaboration mais nous nous concentrons dans la suite de la fiche sur OSIRIS Core qui est la seule version publiée en mai 2023[4].
  - L'ensemble minimal de données est composé de **67** éléments **cliniques** et **65** éléments **omiques**.
- **Organisme en charge** : INCa et Unicancer[5].

2) Voir : Institut National du Cancer
3) Voir l'article de Guérin J, Laizet Y, Le Texier V, Chanas L, Rance B, Koeppel F, Lion F, Gourgou S, Martin AL, Tejeda M, Toulmonde M, Cox S, Hess E, Rousseau-Tsangaris M, Jouhet V, Saintigny P. « OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology ». JCO Clin Cancer Inform. 2021 Mar : OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology
4) Il existe 2 autres versions d'OSIRIS en cours d'élaboration : OSIRIS RWE et OSIRIS Lung (projet en cours, voir : Le SIRIC BRIO postule pour une 3ème labellisation)
5) Voir : Unicancer

APPLICATION

- **Domaine d'application en santé** : oncologie.
- **Principaux cas d'usages** : **recherche clinique** sur le cancer, **essais cliniques**, et en particulier dans le cadre du développement de la **médecine de précision** qui a conduit les SIRICs à mener des essais cliniques de profilage moléculaire [3].
- **Illustration concrète, exemple d'utilisation sur un cas simple** :
  - Un test d'implémentation a été mené sur les données de **300 patients** inclus dans 6 essais cliniques [3] :

→ Dans les données cliniques :

- Une forte hétérogénéité entre les éléments communs de données (Common Data Elements) utilisés dans chaque essai clinique : certains concepts cliniques de l'ensemble minimal de données sont bien représentés (Patient, BiologicalSample, TumorPathologyEvent, Treatment, AdverseEvent, Drug) alors que d'autres sont très peu utilisés (FamilyCancer History, RelatedPathology).
- Des différences notables entre les terminologies utilisées : certains médecins utilisent des terminologies nationales (exemple : la codification de l'Association pour le Développement de l'Informatique en Cytologie et Anatomie Pathologique (ADICAP)) tandis que d'autres utilisent des terminologies internationales (exemple : ICD-O-3, International Classification of Diseases for Oncology).
- Le processus de traitement de données suivant a été adopté : traduction des terminologies nationales vers les terminologies internationales, évaluation de la qualité des données, correction des erreurs et complétion des données manquantes.

→ Dans les données omiques :

- Tous les essais cliniques utilisent le séquençage de nouvelle génération (NGS : Next-generation sequencing) [3], ce qui permet une cohérence des concepts génomiques.
- Données génomiques hétérogènes en raison de la variété des technologies NGS et des outils d'analyse (description des altérations génomiques hétérogènes ou partielles).
- Variabilité des métadonnées (interprétation clinique des variants, etc.).
  - Pour permettre l'interopérabilité de l'ensemble minimal de données OSIRIS, il a été rendu compatible avec le standard HL7 FHIR : la partie génomique du schéma de données a été mappée aux ressources génomiques HL7 FHIR (voir l'élément « Description » plus haut).

DONNÉES

- **Typologie de données concernées** :
  - L'ensemble minimal de données consiste en une liste de **132 éléments communs** de données (Common Data Element), dont 67 correspondent à des données cliniques et 65 à des données omiques[6]. Les **variables** (ou Data Element ou DE, on parle aussi de « concepts ») peuvent correspondre aux **types** suivants : chaîne de caractères, date, nombre entier, nombre décimal.
  - Les concepts représentés dans ces données sont les suivants :
    - **Concepts de données cliniques**[7] : description du patient, consentement du patient, pathologies associées (autres que le cancer ; ex : diabète), antécédents carcinologiques du patient ou de ses apparentés, événement tumoral, analyse survenue au cours d'un événement tumoral, résultat pour un marqueur donné dans le cadre d'une analyse de biologie moléculaire survenue au cours d'un événement tumoral, traitement (chirurgie, chimiothérapie, radiothérapie, immunothérapie), événement indésirable, molécule administrée, échantillon biologique.
    - **Concepts de données omiques** : technologie utilisée, panel, analyse omique, altération génomique, méthode de validation de l'altération, variant, etc.
- **Type de granularité** :
  - La granularité est au niveau d'un **événement carcinologique** : chaque variable est liée à un concept TumorPathologyEvent (TPE), c'est-à-dire à une tumeur primaire, récidive locale ou récidive métastatique[3]. Cela permet de suivre l'évolution de la maladie dans le temps : OSIRIS est un modèle de données temporel fondé sur les événements (event-based temporal model).
- **Utilisation dans plusieurs langues** : le schéma de données et l'ensemble minimal de données sont définis en anglais.

DISPONIBILITÉ DE LA DOCUMENTATION D'IMPLÉMENTATION

- Documentation décrivant la spécification du modèle disponible sur GitHub[8].
- Documentation sur les variables (ou Data Elements) dans les Data Supplements[6].
- Première publication OSIRIS décrivant l'initiative et le schéma de données[3].

6) Voir la documentation des Data Supplements 1, 2 et 3 pour une description précise des éléments de données et de l'ensemble des valeurs qu'ils peuvent prendre : https://ascopubs.org/doi/suppl/10.1200/CCI.20.00094

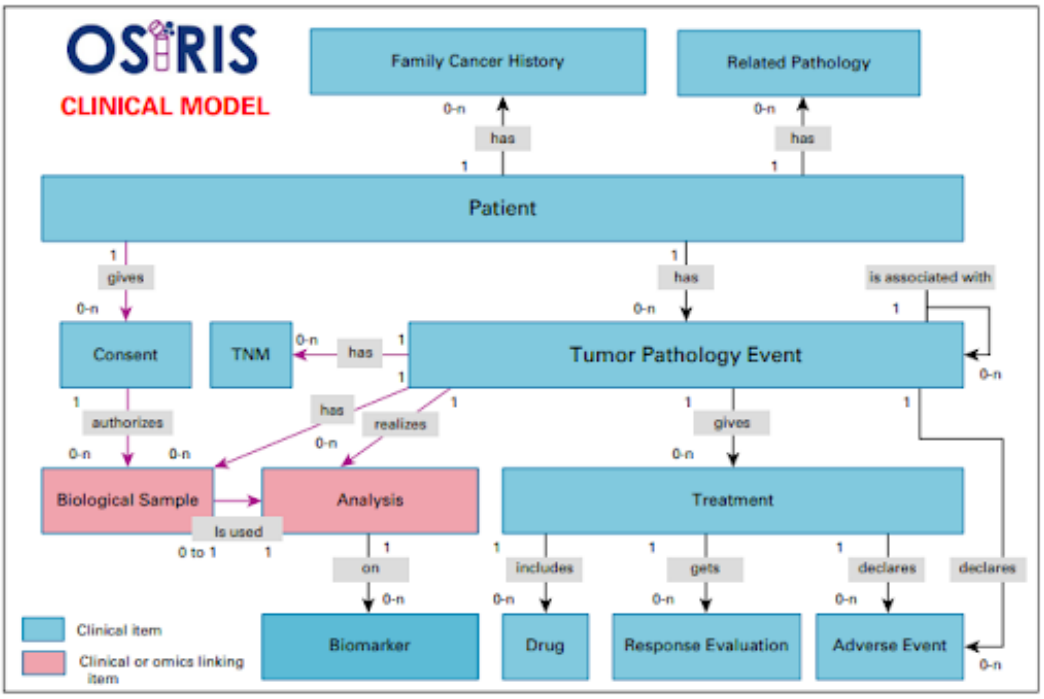
7) Voir la Description du Modèle de Données Cliniques (version 1.0, 2021) : https://github.com/siric-osiris/OSIRIS/blob/master/documentation/OSIRIS_Sp%C3%A9cifications_Mod%C3%A8le_Donn%C3%A9es_Cliniques_TC_v1.0_(DRAFT).pdf

8) Voir le GitHub : https://github.com/siric-osiris/OSIRIS/tree/master/documentation

DESCRIPTION TECHNIQUE DU SCHÉMA DE DONNÉES

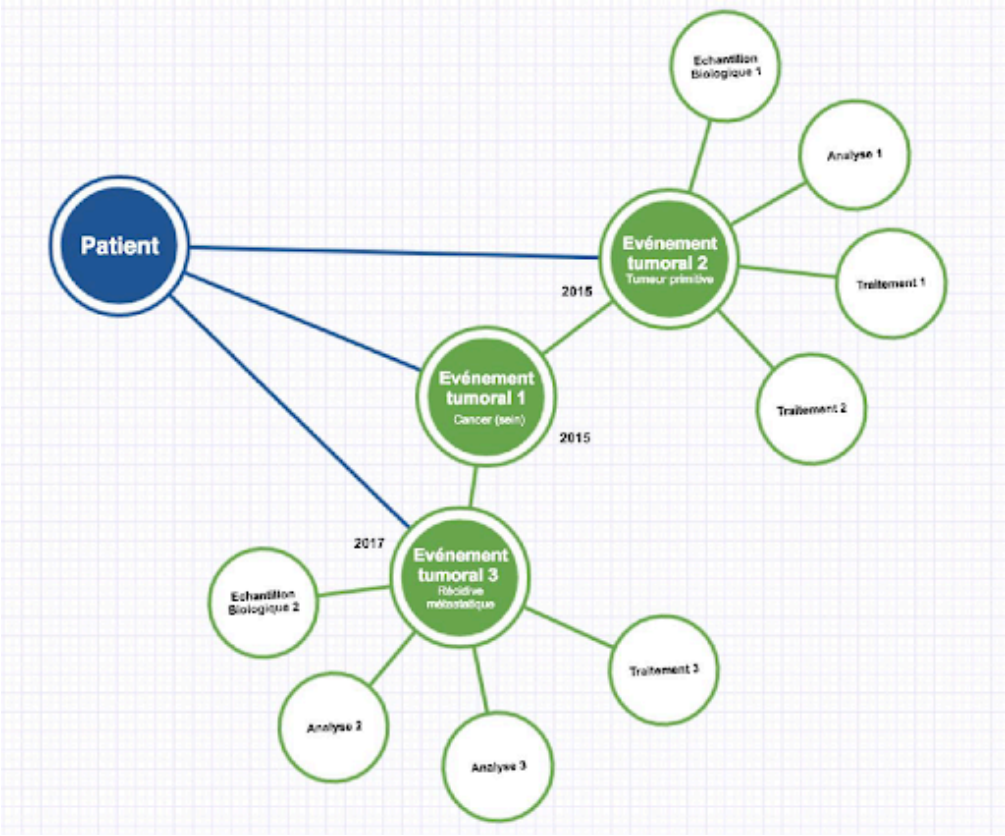
- Le schéma de données OSIRIS est composé (1) d'un **modèle de données cliniques** et (2) d'un **modèle de données omiques**.
  - Dans le **modèle de données cliniques** (voir Figure 1) chaque élément de donnée est lié à un concept de TumorPathologyEvent (tumeur primaire, récidive locale ou récidive métastatique). Le modèle a pour objectif de proposer une **représentation événementielle** de **la maladie carcinologique** de chaque patient du point de vue de ses données cliniques [7](voir Figure 2) :
    - Chaque patient est associé à des **antécédents familiaux** de cancer ainsi qu'à des **pathologies liées.**
  - Chaque patient a un ou plusieurs événements carcinologiques.
    - Pour chaque événement carcinologique, on associe :
      - D'une part les **données cliniques**, à savoir :
        - Le **traitement**.
        - La **réponse** au traitement.
        - Les **effets indésirables** à la suite du traitement.
      - D'autre part les **données omiques**, à savoir : les analyses réalisées sur un échantillon (image, omique, biologie, examen pathologique, etc.). Ce dernier bloc de données fait le lien avec le modèle de données omiques.
  - Le modèle couvre la description des **tumeurs solides** et des changements mineurs seront nécessaires pour couvrir les tumeurs hématologiques (tumeurs développées à partir de cellules du sang)[3].

Figure 1 : Modèle de données cliniques (OSIRIS)



Source : Guérin et al. (2021)

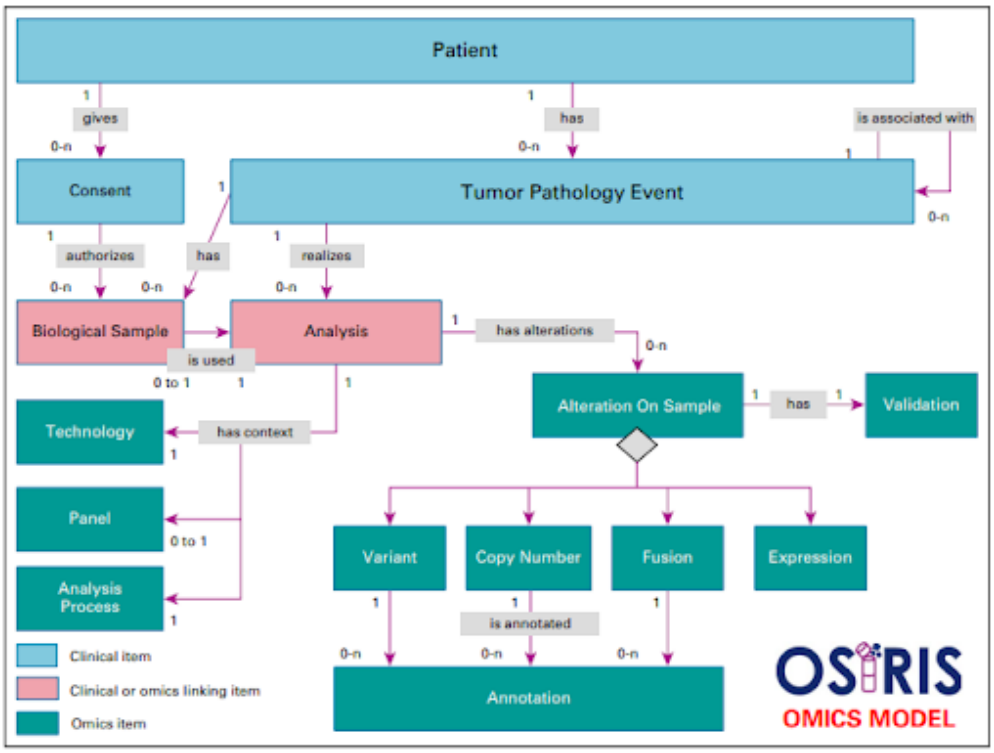
Figure 2 : Exemple des interactions entre les concepts du modèle clinique



Source : Description du Modèle de Données Cliniques (version 1.0, 2021)

- Dans le modèle de données omiques (voir Figure 3), tous les concepts sont liés au TumorPathologyEvent. Les concepts du modèle sont les suivants :
  - Définition du contexte de l'analyse : technologie de séquençage, paramètres de l'analyse, etc.
  - Niveau de confiance : des prédictions (validation) ou de l'annotation des variantes (annotation).
  - Différents types d'altération sont pris en compte dans le modèle : nombre de copies, fusions, expression de gènes, mutations somatiques. La flexibilité du modèle permet d'ajouter de nouvelles données omiques dans le modèle (exemple : épigénétique, protéomique).

Figure 3 : Modèle de données omiques (OSIRIS)



Source : Guérin et al. (2021)

- Les extensions du modèle contiennent d'autres composantes (voir l'élément « Flexibilité du standard, personnalisation » en partie 3. Technique et l'élément « Existence d'extensions certifiées » en partie 5. Utilisation).

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

- **Note : 0,6/1**
- Cette note combine plusieurs sous-critères.
- Elle s'explique par :
  - Le fait que les terminologies ne sont pas imposées, et en particulier il n'y a pas de terminologies locales obligatoires (voir l'élément « Flexibilité dans les choix des terminologies » en partie 3. Technique).
  - L'absence de possibilité de personnalisation du standard malgré l'existence de quelques extensions (voir les éléments « Flexibilité du standard, personnalisation » en partie 3. Technique et « Existence d'extensions certifiées » en partie 5. Utilisation).
  - L'absence de contraintes d'implémentation (voir l'élément « Contraintes d'implémentation » en partie 3. Technique).
  - Une couverture moyenne des cas d'usage (voir l'élément « Principaux cas d'usage » en partie 1. Général).
  - Une faible couverture des domaines d'application (voir l'élément « Domaine d'application en santé » en partie 1. Général).

- La facilité de remplissage des champs du standard dépend :
  - De la **disponibilité des données sources** :
    - Les tables du modèle OSIRIS font référence à des données spécifiques que l'on peut retrouver dans les bases de données des hôpitaux, des laboratoires, et d'autres établissements ayant une expertise en oncologie. Toutefois, en pratique plusieurs difficultés peuvent être rencontrées pour remplir ces champs :
      - Dans les établissements de santé, ces données sont rarement stockées de manière centralisée, elles sont dispersées dans de nombreuses bases de données.
      - Selon les pratiques de l'établissement et des professionnels de santé, le niveau de complétude des champs ainsi que la profondeur de l'historique varient.
      - Les établissements de santé ont en général une vision limitée du parcours de soins, qui se limite au périmètre de leur établissement.
  - De la facilité à réaliser le **mapping des données sources** :
    - Le modèle OSIRIS est extensible, modulaire et dispose d'extensions permettant de traiter divers types de données (imagerie, radiothérapie, etc., voir l'élément « Flexibilité du standard, personnalisation » en partie 3. Technique et l'élément « Existence d'extensions certifiées » en partie 5. Utilisation).

2. Gouvernance

LIBRE ACCÈS AUX SCHÉMAS DE DONNÉES : OUI.

- Liste des concepts cliniques et omiques disponible sur GitHub [9] et la documentation associée est disponible également [6].

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

- Le modèle OSIRIS est en accès libre sur GitHub et n'est pas protégé par une licence.[10]

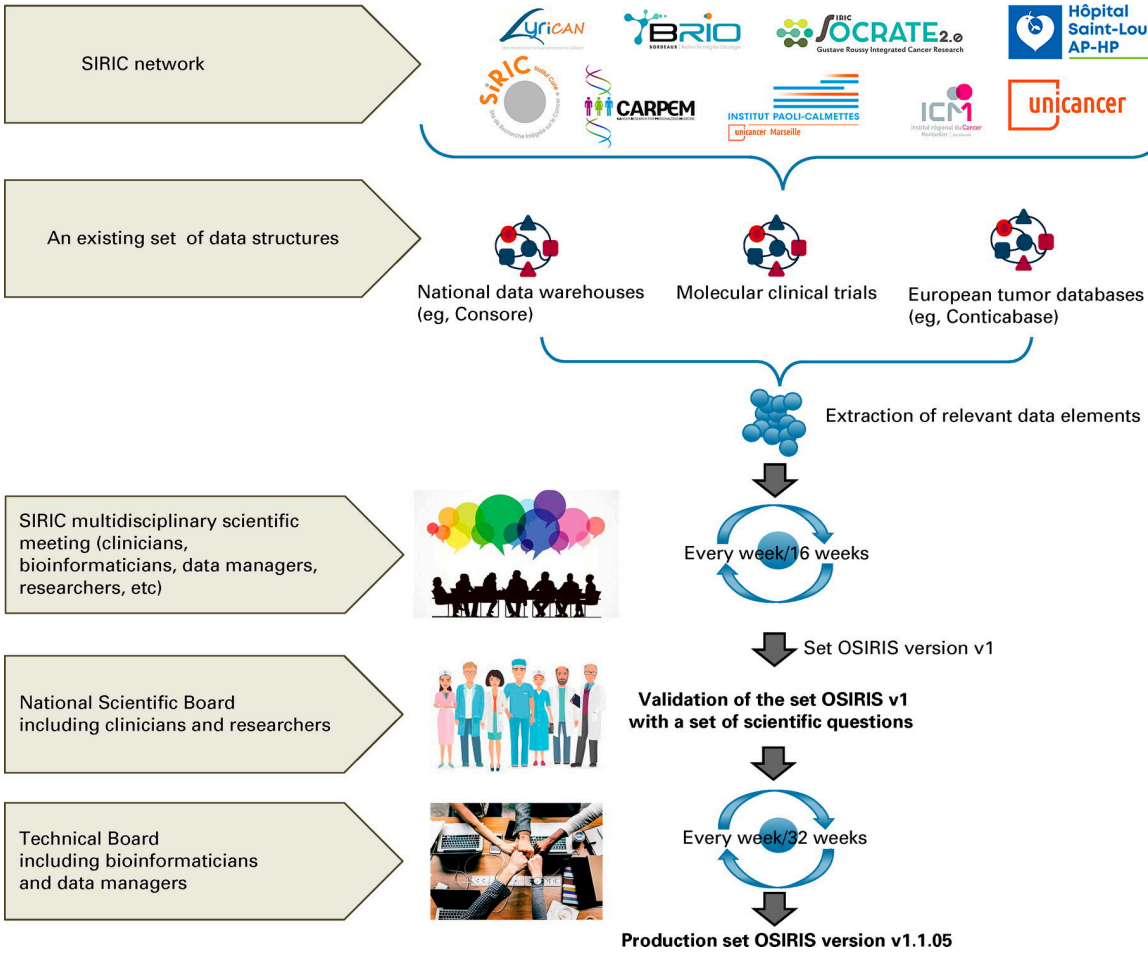
PROCESSUS DE PRISE DE DÉCISION SUR LE STANDARD :

- Le **groupe de travail**, mené par l'INCa, et contributeur au modèle de données initial est composé d'Unicancer, de plusieurs Centres de Lutte Contre le Cancer (CLCC) [11], de Centres Hospitaliers Universitaires (CHU) [12] et des 8 SIRICs [13]. Les principaux CLCC ayant contribué sont l'Institut Curie, l'Institut Bergonié et le Centre Léon Bérard.
- La méthodologie d'élaboration du modèle OSIRIS a consisté à organiser des réunions hebdomadaires de plusieurs groupes nationaux (voir Figure 4) :
  - Groupe SIRIC multidisciplinaire (composé d'oncologues, spécialistes d'informatique médicale, bio-informaticiens, épidémiologistes, bio-statisticiens, data managers, chercheurs, etc.) : sélection des données les plus pertinentes à inclure dans le modèle (Data Elements).
  - Groupe scientifique (aussi appelé National Scientific Board, composé d'oncologues, de chercheurs en recherche translationnelle en oncologie et de data protection officers) : examen de la conformité des données sélectionnées précédemment (Data Elements) au Règlement Général sur la Protection des Données (RGPD) et à la réglementation de la Commission Nationale de l'Informatique et des Libertés (CNIL).

9) Voir le Github.
10) Voir le troisième document Data Supplement qui compare notamment l'existence de licences associées à OSIRIS mCODE et OMOP-CDM. Pour OSIRIS, il est indiqué : « Licensing : None ».
11) Principalement l'Institut Curie, l'Institut Bergonié, le Centre Léon Bérard, l'Institut Gustave Roussy, l'Institut du cancer de Montpellier, l'Institut Paoli-Calmettes.
12) Principalement le CHU de Bordeaux, l'Hôpital Européen Georges Pompidou et l'Hôpital Saint-Louis de l'AP-HP.
13) Les SIRICs : CARPEM, CURAMUS, Curie, EpiCURE, ILIAD, InSiTu, LYriCAN, Montpellier Cancer.

- Groupe technique (aussi appelé Technical Board, composé de bio-informaticiens et data managers) : identification des terminologies nationales et internationales les plus pertinentes (exemple : la Classification Commune des Actes Médicaux (CCAM) pour les actes médicaux) et mapping des données (Data Elements) avec ces terminologies.

Figure 4 : Méthodologie d'élaboration de la première version d'OSIRIS



Source : Guérin et al. (2021)

- La procédure de contribution aux concepts (nouvelles extensions ou mises à jour de concepts existants) est décrite dans le GitHub [14] : la procédure est libre, les utilisateurs peuvent proposer des modifications (via une pull request) qui seront examinées par le groupe OSIRIS.
- Des suggestions de modifications ou des bogues peuvent également être signalés dans le repository GitHub dédié [15].
- Des questions peuvent être envoyées à l'adresse contact@siric-osiris.fr [14].

14) Voir la procédure de contribution.
15) Voir le Github.

MATURITÉ DU STANDARD

- Fréquence de mise à jour :
  - La première version du schéma de données OSIRIS (v1) est sortie en mai 2018.
  - Depuis cette date, 3 mises à jour de la v1 ont été publiées (v1.1.03, v1.1.04 et v1.1.05)[16] à la suite des échanges avec les groupes scientifique et technique (voir l'élément « Processus de prise de décision sur le standard » ci-dessus).
  - En mai 2023, la dernière version est la **v1.1.05** sortie en février 2019.
- Maturité :
  - **Note : 0,4/1.**
  - Cette note combine plusieurs sous-critères. Elle s'explique par :
    1. La publication d'une version stable.
    2. L'absence à date de mise à jour de la version principale.
    3. Le jeune âge du standard.
    4. Son utilisation dans le monde restreinte à l'échelle d'institutions (voir l'élément « Adoption du standard » en partie 4. Valorisation).
    5. L'absence d'adoption officielle par un ou plusieurs pays ou par une organisation de référence (voir l'élément « Adoption du standard » en partie 4. Valorisation).

EXISTENCE DE FINANCEMENTS POUR STANDARDISATION : NON.

16) Voir la liste des versions.

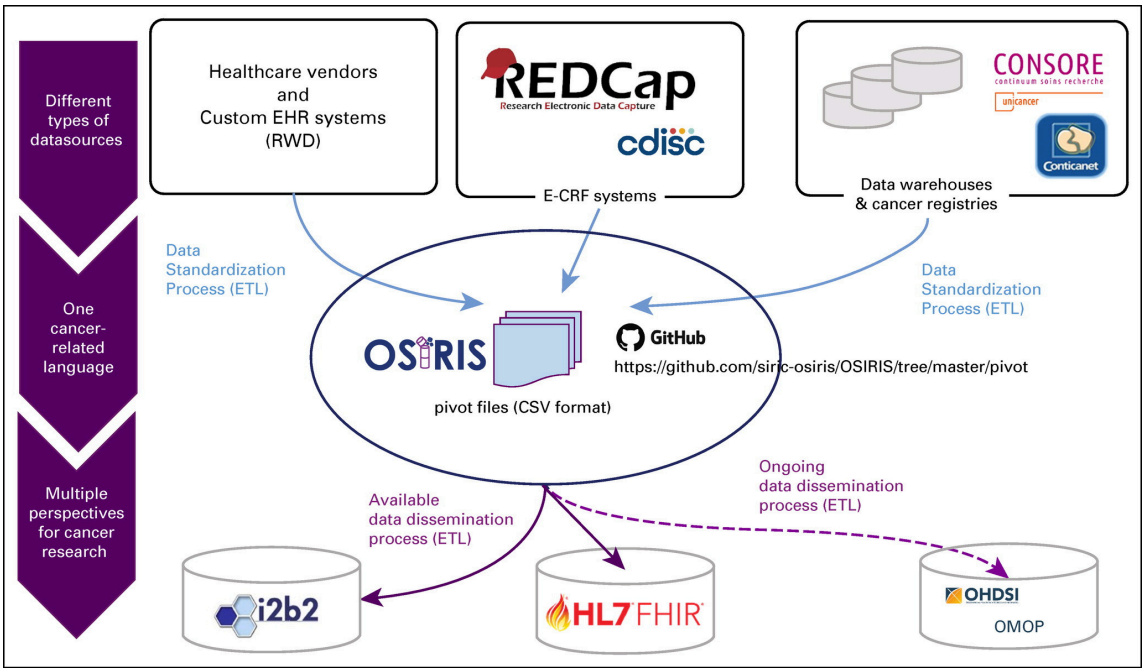
3. Technique

CAPACITÉ DE TRADUCTION VERS UN AUTRE STANDARD (INTRA TYPES DE STANDARDS)

- **OSIRIS** est conçu comme un modèle de données dont l'un des objectifs principaux est d'être interopérable. Il définit l'ensemble minimal de données théoriques, issu des données source (EHR, systèmes e-CRF tels que REDCap ou cdisc, ou entrepôt de données du type ConSoRe), permettant d'analyser l'évolution de la maladie carcinologique. Il ne définit que la structure théorique, et non la structure physique des bases de données. Pour cela, des standards internationaux, qui contiennent une implémentation physique, sont utilisés dans une étape complémentaire, selon l'objectif recherché (voir Figure 5 ci-dessous), et les processus ETL associés sont développés en partenariat avec des sociétés tierces :
    - **Processus ETL OSIRIS-FHIR** (travail réalisé en partenariat avec la société Arkhn)[17] : les Data Elements qui constituent l'ensemble minimal de données ont été associés aux ressources FHIR correspondantes (voir Figure 5 ci-dessous).
    - **Processus ETL OSIRIS-i2b2** [18] : pour rendre compatible OSIRIS et i2b2, un guide détaille les étapes de déploiement de l'instance i2b2/SHRINE et contient les fichiers permettant de charger les Data Elements OSIRIS.
    - Le consortium OSIRIS travaille actuellement à développer le processus ETL OSIRIS – OMOP-CDM[3] (voir Figure 5 ci-dessous).
    - Le modèle OSIRIS a été comparé aux modèles mCODE et OMOP-CDM dans le Data Supplement n°3 (sur les caractéristiques générales, les composantes du modèle, les concepts disponibles et les terminologies utilisées)[6].

17) Voir le guide d'implémentation FHIR d'OSIRIS.
18) Voir le guide d'installation Docker pour déployer l'instance i2b2/SHRINE.

Figure 5 : Schéma d'intercommunication d'OSIRIS avec des sources de données et avec d'autres standards internationaux



Source : Guérin et al. (2021)

COMMUNICATION AVEC D'AUTRES STANDARDS (INTER TYPOLOGIES DE STANDARDS)

OSIRIS utilise plusieurs terminologies nationales et internationales, dont la CIM-10 pour les maladies, sa variante dédiée à l'oncologie (CIM-O-3), la CCAM pour les actes médicaux, LOINC pour les concepts génomiques (voir Tableau 1 ci-dessous).

Tableau 1 : Principales terminologies utilisées dans OSIRIS :

Data Domain	National and International Ontologies and Terminologies
Patient characteristics	Fast Healthcare Interoperability Resources (FHIR, 3rd edition) Unified Medical Language System (UMLS) WHO classification (performance status)
Disease characteristics	International Classification of Disease for Oncology (ICD-O-3, 3rd edition) International Statistical Classification of Diseases and Related Health Problems (ICD, 10th edition) UICC TNM Classification of Malignant Tumors
Drug	Anatomical Therapeutic Chemical Classification System (ATC, 5th level)
Adverse events	Common Terminology Criteria for Adverse Events (CTCAE, 5th edition)
Response evaluation	RECIST, version 1.1
Medical act	Classification of the French Social Security (National Health Service)
Genomic concepts	Logical Observation Identifiers Names and Codes (LOINC) HL7 Fast Healthcare Interoperability Resources (HL7 FHIR)

Source : Guérin et al. (2021)

Flexibilité dans les choix des terminologies :

Le modèle OSIRIS utilise les terminologies nationales et internationales les plus pertinentes dans chaque domaine pour assurer l'interopérabilité. Lorsqu'il n'existe pas de définition standard, le groupe de travail a créé sa propre terminologie ad hoc (voir la description du travail du groupe technique dans l'élément « Processus de prise de décision sur le standard » en partie 2. Gouvernance, le Tableau 1 ci-dessus pour la liste des principales terminologies utilisées et le Data Supplement n°2 pour la liste exhaustive de l'ensemble des terminologies du modèle[6]).

Flexibilité du standard, personnalisation :

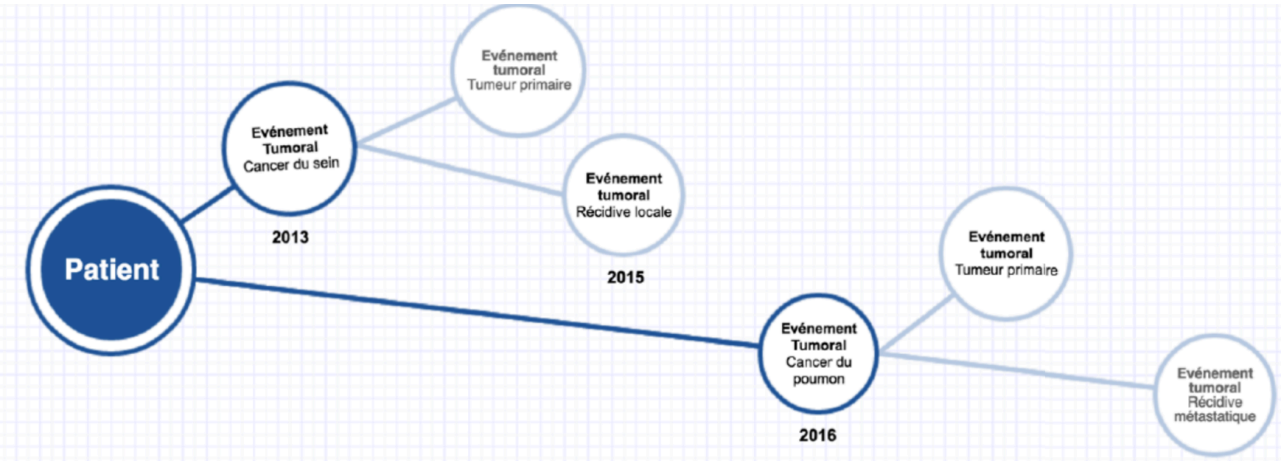
Dans le modèle, conçu pour être **modulaire** et **extensible**, il est également possible d'ajouter des données **biologiques** ou **omiques** supplémentaires (exemple : protéomiques), ainsi que d'autres types de données. Les extensions existantes permettent de traiter les données d'imagerie, les données radiomiques et les données de radiothérapie (voir l'élément « Existence d'extensions certifiées » en partie 5. Utilisation).

Performance :

- **Note : 1/1** en raison du faible nombre de jointures nécessaires pour réaliser des requêtes. En effet, le niveau de normalisation du modèle étant moyen, le nombre de jointures nécessaires pour chaque requête est restreint.
- L'implémentation native d'OSIRIS utilise le schéma construit autour de la table d'évènements carcinologiques (**TumorPathologyEvent**)[19]. Cette table contient le lien circulaire vers elle-même qui permet de gérer la hiérarchie des événements.
  - Par exemple, un **événement carcinologique « parent »** (ex : cancer du sein) peut être associé à plusieurs **autres événements** (ex : tumeur primaire, récidive locale, récidive métastatique, voir Figure 6). Dans ce cas-là, l'identifiant de l'événement « parent » sera contenu dans le champ « ParentRef » et les identifiants des événements « enfants » seront contenus dans les champs InstanceId (voir Figure 7).
  - Cependant, cette **structure hiérarchique** implique d'utiliser soit des **requêtes récursives** soit une **série de requêtes consécutives** pour analyser l'évolution de ces événements dans le temps ou d'autres indicateurs qui nécessitent de prendre en compte la relation « parent » - « enfant ».
- Le modèle centré sur les événements d'OSIRIS impose également des **restrictions sur la nature de l'organisation des clés étrangères**. Au lieu d'une clé simple, il utilise une clé composée de 2 à 4 champs, ce qui impose un nombre minimum de tables impliquées dans l'analyse et, par conséquent, un nombre minimum de jointures.

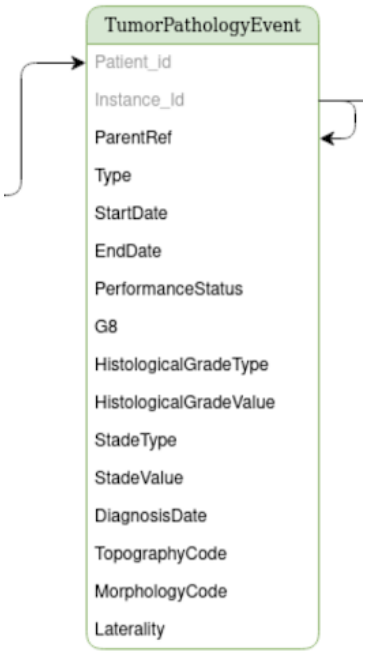
19) Voir https://github.com/siric-osiris/OSIRIS/blob/master/documentation/MPD_OSIRIS_model_v1.1.05.png

Figure 6 : Exemple de relations « parent » - « enfant » entre les événements carcinologiques



Source : Description du Modèle de Données Cliniques (version 1.0, 2021)

Figure 7 : Extrait du schéma de données OSIRIS (table TumorPathologyEvent)



Source : Description du Modèle de Données Cliniques (version 1.0, 2021)

COMPLEXITÉ DU MODÈLE

- **Note : 0,4/1**
- Cette note combine plusieurs sous-critères. Elle s'explique par :
  1. Les nombreuses tables avec des liens entrants et sortants.
  2. Le niveau moyen de normalisation du modèle.
  3. L'absence de tables larges.
- Le modèle comporte **18 tables, 213 champs** (dont 132 champs cliniques et omiques et 81 champs techniques) et **24 relations**.
- Le modèle utilise le type d'organisation des tables « **flocon de neige** » où certaines tables ne sont pas directement liées à la table d'événement. Cela permet de réduire la redondance d'information en utilisant des relations entre les différentes tables.
- Cependant, toutes les tables ne sont pas normalisées de cette manière : certaines contiennent des informations redondantes au lieu d'utiliser une table de dimension supplémentaire. Par conséquent, le niveau de normalisation n'est pas très élevé.
  - Par exemple, les tables **Variant**, **CopyNumber**, **Fusion** et **Expression** contiennent une partie **Validation** qui inclut les champs Type, Method et Status pour chaque clé PatientId et Analysis_Ref.
  - Pour normaliser ces tables, on pourrait utiliser une table supplémentaire Validation qui contiendrait une clé primaire Validation_Id ainsi que l'ensemble des valeurs pouvant être prises par les champs Type, Method et Status.
- Ainsi, la **complexité du modèle OSIRIS est réduite** comparativement à un schéma « flocon de neige » complètement normalisé : on observe un équilibre entre l'augmentation de relations entre les tables et la redondance d'information.

CONTRAINTES D'IMPLÉMENTATION

OSIRIS n'impose pas de contrainte d'implémentation.

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

- Le standard **ne définit pas la technologie de stockage**. Il est possible d'utiliser tous les types de technologies de stockage pour les données structurées (y compris SGBDs, fichiers délimités, etc.).
- À titre d'exemple, dans l'article de Guérin et al. (2021)[3], le format CSV est utilisé dans les fichiers pivot pour stocker l'information provenant des établissements de santé dans un format correspondant au schéma OSIRIS avant de la traduire dans un schéma international cible (i2b2, OMOP-CDM, HL7 FHIR, etc.).

TYPE DE TECHNOLOGIE DE REQUÊTAGE

Le standard **ne définit pas les technologies de requêtage**, elle dépend de l'implémentation.

NEUTRALITÉ TECHNOLOGIQUE

OSIRIS est **neutre** technologiquement. Il définit **le schéma de données** mais il est indépendant de la technologie utilisée.

INTENSITÉ DE LA PERTE DE DONNÉES AU MAPPING

Le schéma de données OSIRIS a été conçu pour être utilisé comme un **modèle commun** de l'analyse de la maladie carcinologique, dans une **étape intermédiaire** entre la lecture des données sources et l'implémentation d'un standard international cible[3]. L'utilisation du schéma OSIRIS implique donc **2 étapes de mapping** (voir l'élément « Capacité de traduction vers un autre standard (intra types de standards) » et la Figure 5 en partie 3. Technique) successives :

- 1) Depuis les systèmes sources de stockage de données dans les établissements de santé vers le langage commun OSIRIS. Au sujet de cette étape de mapping, en date de mai 2023, il n'y a pas eu d'évaluation quantitative de l'intensité de la perte de données.
- 2) Depuis le langage commun OSIRIS vers les standards internationaux d'interopérabilité cibles. Pour cette seconde étape, l'intensité de la perte de données dépend du standard cible choisi et de ses caractéristiques.

COMPÉTENCES TECHNIQUES ET MÉTIER NÉCESSAIRES POUR UTILISER LE STANDARD

- **Profil Data Engineer/Database Administrator** pour mettre en place le modèle de données physique pour la technologie de stockage.
- **Profil Data Engineer** pour développer des flux de données transformant les schémas source vers OSIRIS puis vers le schéma cible (i2b2, OMOP-CDM, etc.)
- **Profil informaticien médical ou bio-informaticien** pour définir le mapping entre les systèmes source et OSIRIS.

4. Valorisation

ACCESSIBILITÉ À DES RESSOURCES DE FORMATION

En mai 2023, il n'existe pas de ressources de formation (uniquement des ressources de documentation du modèle).

DISPONIBILITÉ DE LA DOCUMENTATION SCIENTIFIQUE DÉMONTRANT L'INTÉRÊT

- L'**Institut Curie** liste l'ensemble de ses projets de recherche en cours, dont 2 utilisant le modèle OSIRIS[20]:
  - Un projet initié en septembre 2020 (sur environ 110 patients) : étude rétrospective nationale sur les sarcomes des tissus mous chez les enfants, afin de mieux caractériser l'importance du traitement local.
  - Un projet initié en novembre 2021 (sur environ 200 patients, étude également menée en parallèle au CHU de Bordeaux) : étude des facteurs de modulation de la réponse au traitement pour des patients atteints d'un cancer du poumon métastatique et ayant bénéficié d'une thérapie ciblée sur la base d'une anomalie des gènes EGFR et/ou ALK.

ADOPTION DU STANDARD

- Adoption officielle : en mai 2023, il n'y a pas encore eu d'adoption officielle du modèle OSIRIS.
- Utilisation sur le marché :
  - Les institutions **membres de l'initiative OSIRIS** utilisent déjà le schéma de données (Institut Curie, Institut Bergonié, Centre Léon Bérard, Institut du Cancer de Montpellier, Institut Paoli-Calmettes, Institut Gustave Roussy, CHU de Bordeaux, Hôpital Européen Georges Pompidou, Hôpital Saint Louis, Unicancer)[21]. À terme, l'objectif est de créer un large réseau de bases de données fédérées.

FOURNISSEURS DE SERVICE AYANT L'EXPERTISE EN FRANCE

En mai 2023, il n'y a pas de fournisseurs de service ayant l'expertise en France.

QUALITÉ DES DONNÉES

- **Existence d'un label de qualité** : non.
- **Outils de vérification de la qualité des données** : non.

20) Voir le site de l'Institut Curie.
21) Voir OSIRIS : a national data sharing project - www.en.ecancer.fr.

5. Utilisation

SIMPLICITÉ D'USAGE

- **Note : 0,3/1**
- Cette note combine plusieurs sous-critères. Elle s'explique par :
  1. L'absence d'accès à des ressources officielles de formation (voir l'élément « Accessibilité à des ressources de formation » en partie 4. Valorisation)
  2. La lisibilité du schéma par un humain (voir l'élément « Lisible par un humain » en partie 5. Utilisation)
  3. L'absence d'outils de gestion de la qualité des données (voir l'élément « Qualité des données » en partie 4. Valorisation)
  4. Le nombre élevé 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)
- OSIRIS étant un **modèle de données théorique** qui définit l'**ensemble minimal de données** à considérer, il est **simple** à utiliser. L'implémentation du modèle dans un standard international (ex : i2b2, FHIR ou OMOP-CDM) peut cependant s'avérer plus compliquée, en fonction des caractéristiques du standard choisi.

EXISTENCE D'UNE COMMUNAUTÉ EN LIGNE ET DEGRÉ D'ACTIVITÉ : NON.

OUTILS DE MAPPING : NON.

OUTILS COMPATIBLES : NON.

DÉCRIRE LES ÉTAPES NÉCESSAIRES POUR LA STANDARDISATION

OSIRIS est un langage commun pour l'analyse de la maladie carcinologique. Il est utilisé dans une étape intermédiaire de traduction des données, entre l'import des données sources et la standardisation dans un format international (voir la Figure 5).

Étape 1 : processus de standardisation des données sources (construction d'un processus ETL)

- Les données sources peuvent exister dans différents formats : EHR, système eCRF (du type REDCap ou cdisc), entrepôts de données (exemple : l'entrepôt ConSoRe[22] pour les Centres de Lutte Contre le Cancer (CLCC) en France), registre du cancer.
- Le processus ETL pour traduire ces données dans le format OSIRIS n'existe pas mais les règles à adopter sont définies dans la spécification et dans les fichiers pivot disponibles sur le GitHub au format CSV[23]. Un processus ETL peut donc être établi dans chaque institution pour traduire les données sources dans le langage OSIRIS.

Étape 2 : processus de dissémination des données (construction d'un processus ETL)

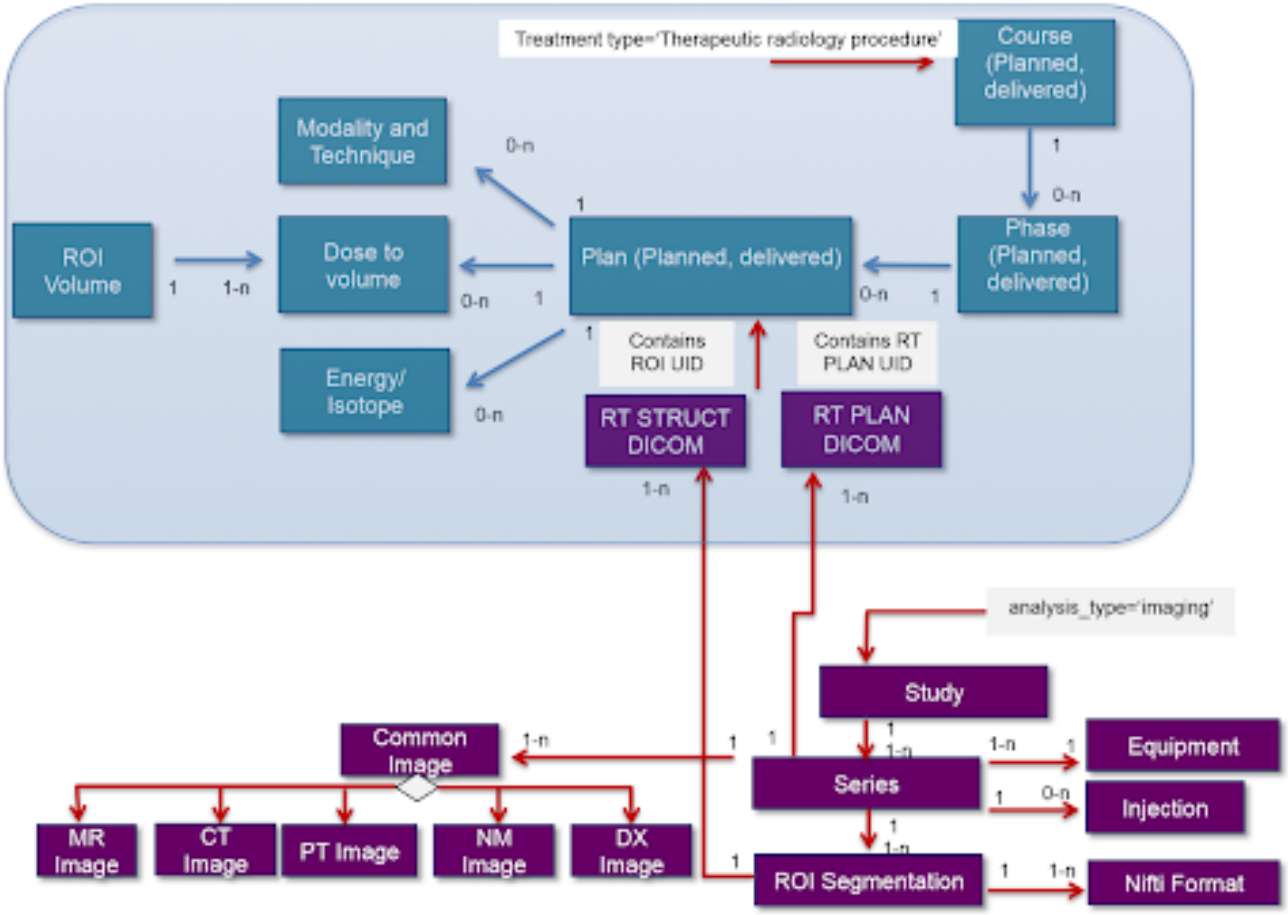
- Les données traduites dans le langage OSIRIS sont ensuite rendues compatibles avec un ou plusieurs standards internationaux (i2b2, HL7 FHIR, OMOP-CDM), en fonction de l'objectif recherché, de manière à assurer l'interopérabilité.

EXISTENCE D'EXTENSIONS CERTIFIÉES

Il existe **plusieurs extensions du modèle OSIRIS** : il s'agit des modèles d'imagerie, de radiomique et de radiothérapie (voir Figure 8 ci-dessous pour la radiothérapie), qui permettent de traiter et d'intégrer les données spécifiques concernées[24].

22) Voir la documentation sur l'entrepôt ConSoRe.
23) Voir le Github.
24) Voir la liste des extensions.

Figure 8 : Modèle de données de l'extension radiothérapie



Source : Github

BIBLIOTHÈQUE DE REQUÊTES TYPES : NON.

LISIBLE PAR UN HUMAIN : OUI.

Le modèle OSIRIS se présente sous forme tabulaire et les noms des concepts et des variables sont compréhensibles et renseignent directement sur le contenu. De plus, la succession des événements est également intelligible grâce à la structure longitudinale du schéma.

6. Glossaire des acronymes

1. **ADICAP :** Association pour le Développement de l'Informatique en Cytologie et Anatomie Pathologique
2. **ADN :** Acide DésoxyriboNucléique
3. **ALK :** Anaplastic Lymphoma Kinase
4. **AP-HP :** Assistance Publique – Hôpitaux de Paris
5. **API :** Application Programming Interface
6. **ARN :** Acide RiboNucléïque
7. **CCAM :** Classification Commune des Actes Médicaux
8. **CHU :** Centre Hospitalier Universitaire
9. **CIM :** Classification Internationale des Maladies (ICD en anglais)
10. **CLCC :** Centre de Lutte Contre le Cancer
11. **CNIL :** Commission Internationale de l'Informatique et des Libertés
12. **CSV :** Comma Separated Values
13. **DBA :** DataBase Administrator
14. **DE :** Data Element
15. **DQD :** Data Quality Dashboard
16. **DWH :** Data WareHouse
17. **eCRF :** electronic Case Report Form
18. **EGFR :** Epithelial Growth Factor Receptor
19. **EHR :** Electronic Health Record
20. **ETL :** Extract Transform Load
21. **HL7 FHIR :** Heaven Level 7 Fast Health Interoperability Resources
22. **2b2 :** Informatics for Integrating Biology & the Bedside
23. **ICD :** International Classification of Diseases
24. **ICD-O-3 :** ICD for Oncology, 3rd version
25. **INCa :** Institut National du Cancer
26. **LOINC :** Logical Observation Identifiers Names & Codes
27. **NGS :** Next-generation sequencing
28. **OHDSI :** Observational Health Data Sciences and Informatics
29. **OMOP-CDM :** Observational Medical Outcomes Partnership Common Data Model
30. **OSIRIS :** GrOupe inter-SIRIC sur le paRtage et l'Intégration des donnéeS clinico-biologiques en cancérologie
31. **RGPD :** Règlement Général sur la Protection des Données
32. **SGBD :** Système de Gestion de Base de Données
33. **SHRINE :** Shared Health Research Information Network
34. **SIRIC :** SItes de Recherche Intégrés sur le Cancer
35. **TNM :** Tumor Node Metastases
36. **TPE :** TumorPathologyEvent