COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Austin Mast (Florida State University), May 19, 2021

Title: *Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses*

Austin Mast CIC Database Profile

NSF Award #: 2033973

YouTube Recording with Slides

May 2021 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Austin Mast:

Slide 1

Merci beaucoup pour l'invitation. J'aimerais commencer par dire qu'il s'agissait d'un travail d'équipe et que nous disposions d'une excellente équipe.

Des crises comme la pandémie émergent et nous éprouvons un besoin urgent de données. Parfois les données dont nous avons besoin sont sur la biodiversité. Dans ce cas, nous aimerions avoir plus de connaissances sur les chauve-souris. Dans d'autres cas, comme les marées noires, il peut s'agir de l'ensemble du biote d'une région particulière pour lequel nous avons besoin de données. Avant notre travail, nous n'avions pas d'ensemble de protocoles de réponse en cas de crise pour améliorer rapidement les données relatives à une source importante d'informations sur la biodiversité : à savoir les 3 à 4 milliards de spécimens de la biodiversité dans le monde.

Slide 2

Comme nous le voyons pendant la pandémie actuelle, il peut s'agir d'un petit sous-ensemble de spécimens qui deviennent critiques pour la réponse à la crise. Les spécimens sont associés à des informations qui documentent ce qui a été collecté, où cela a été collecté, qui l'a collecté, et d'autres informations. Il existe également des capsules temporelles d'informations potentielles, car les données génomiques peuvent souvent être dérivées du spécimen ou de ses agents pathogènes. Il ne s'agit que de

quelques spécimens de l'espèce de chauve-souris en fer à cheval chez laquelle le parent le plus proche du SRAS-CoV-2 a été découvert. Il s'agit de *Rhinolophus affinis*.

Slide 3

Nous avons ciblé un ensemble de trois familles très proches, qui inclut la famille de la *Rhinolophus affinis* pour l'amélioration des données sur les spécimens.

Slide 4

Il s'agit des cartographies des spécimens de chauve-souris fer à cheval/Rhinolophe dans les deux principaux agrégateurs de données sur les spécimens. Je veux mettre en avant que les données collectées et agrégées sont précieuses dans leur état actuel. Cependant, les données présentent certaines qualités qui peuvent être améliorées par l'examen des données dans leur ensemble et les données ont été créées sur deux décennies ou plus, ce qui signifie que les données n'ont pas toutes bénéficié de notre compréhension actuelle des bonnes pratiques et de la disponibilité de logiciels permettant d'améliorer certaines étapes.

Slide 5

Nous nous sommes concentrés sur l'amélioration des données de ces manières et je vais les passer en revue avec vous en caractères gras. Si vous prêtez attention au changement de titre de la diapositive, vous pourrez suivre notre progression relativement rapide dans ces activités.

Slide 6

Les données sur les spécimens provenant des deux principaux agrégateurs se sont chevauchées, mais elles ne sont pas identiques. La déduplication de leurs enregistrements a permis d'obtenir environ 90 000 enregistrements dans le champ d'application.

Slide 7

Les archives sont conservées par 118 institutions dans le monde entier. Les 10 premières de ces institutions se partagent 63 % des enregistrements.

Slide 8

Nous ne pouvions attribuer ou évaluer les coordonnées des collections - les emplacements des collections - que lorsque ces emplacements étaient décrits dans les données partagées. Et environ deux tiers des enregistrements contenaient ces informations. Parmi ceux-ci, environ deux tiers sont arrivés avec des coordonnées pré-assignées et un tiers n'en avait pas. Nous avons pu évaluer ou attribuer des coordonnées dans 95 % des cas possibles et nous avons modifié des coordonnées préexistantes dans la moitié des cas. Le déplacement médian d'une coordonnée préexistante était de six kilomètres.

Slide 9

Il est important de noter que les champs de métadonnées pertinents sont passés de pratiquement

vides à pratiquement complets grâce à l'ajout d'informations utiles telles que le protocole de géoréférencement et les ressources de géoréférencement.

Slide 10

Dans ce résumé, au niveau des pays, vous pouvez voir où le plus grand nombre de spécimens a été collecté par la taille du diagramme circulaire et le nombre relatif de nouvelles coordonnées ajoutées aux spécimens de ces pays.

Slide 11

Voici les coordonnées des lieux de collecte pour chacune de nos familles focales.

Slide 12

Nous avons comparé nos coordonnées avec d'anciennes cartes de la répartition des espèces quand elles étaient disponibles à l'Union Internationale pour la Conservation de la Nature (UICN). Voici un exemple de circonscription de l'aire de répartition en rouge pour une espèce, il s'agit une nouvelle fois de *Rhinolophus affinis*, et nos coordonnées pour cette espèce en vert. Nous avons découvert que les spécimens géoréférencés suggèrent une extension de l'aire de répartition pour 153 des 169 espèces de chauves-souris pour lesquelles nous disposons de ce genre de cartes. C'est une expansion significative de notre compréhension des endroits où trouver les chauves-souris. Voici une capture d'écran d'un explorateur de données sur les chauves-souris fer à cheval basé sur internet, destiné aux évaluateurs de cartes de l'UICN et à d'autres parties prenantes, qui permet de consulter les coordonnées des localités par rapport aux cartes actuelles de l'UICN, avec des liens vers les enregistrements complets dans notre système.

Slide 13

Les enregistrements sont associés avec 2930 valeurs distinctes qui référencent les personnes qui ont collecté ou identifié les spécimens. Nous avons été capable d'assigner 803 identifiants uniques à un sous-ensemble de ces valeurs. Ces identifiants uniques, ou ORCID IDs quand la personne est vivante, et Wikidata IDs quand la personne est décédée. 437 valeurs additionnelles représentant 359 personnes ont été assignées raisonnablement à des personnes vivantes mais qui n'ont pas encore un ORCID ID.

Slide 14

Pour cela, nous avons embauché 34 personnes qui sont principalement des experts des chauves-souris venant de 13 pays. Ces experts et nos curateurs de données ont découvert qu'ils pouvaient associer à peu près la moitié des enregistrements à un identifiant unique pour les collecteurs de spécimen et environ deux tiers les identificateurs de spécimen.

Slide 15

La valeur de ce travail pour une réponse de crise peut ne pas paraître évidente, mais cela pourrait être

l’une des choses les plus importantes que nous ayons faites. Nous avons identifié 117 personnes avec des ORCID IDs avec de l’expérience dans la collecte de chauves-souris. Vous pourriez objecter qu’il est facile de trouver les experts de chauve-souris - il suffit de se rapprocher de leur organisation ou de faire une recherche bibliographique. En réalité, les collecteurs de chauve-souris et ceux que vous pourriez trouver avec ces deux méthodes ne s'intersectent que partiellement. Les collecteurs de chauve-souris se répartissent en fait parmi une grande variété de professions incluant certaines qui ne sont pas reliées à la biologie. Voici quelques descriptions de collecteurs avec une expérience précieuse qui font du travail de terrain parfois dans des régions reculées. Rappelez-vous, nous avons aussi identifié 359 collecteurs de chauves-souris vivants qui n’ont pas de ORCID ID. L'ensemble constitue un rolodex de contacts potentiels pour ceux d'entre vous qui doivent retourner sur le terrain pour relocaliser des populations de chauves-souris.

Slide 16

Avant l'amélioration des données, 5,5 % des enregistrements contenaient des informations sur les séquences associées. Nous avons identifié environ 1 100 spécimens supplémentaires auxquels nous pouvions associer les nouvelles séquences que nous avons trouvées.

Slide 17

Notre version des données, et nos protocoles, sont partagés sur Zenodo. Nous avons donc pavé un chemin que d’autres peuvent suivre pour améliorer la qualité des données rapidement pendant la prochaine crise. Nous nous attendons à partager notre version finale complète très bientôt. Néanmoins, j’aimerais noter que notre version actuelle est très proche de la version finale. Nous sommes proches de soumettre un manuscrit sur ce travail et nous attendons à mettre largement à disposition l’explorateur de données pour les chauves-souris fer à cheval pour ceux qui en ont besoin.

Slide 18

L’UE a récemment annoncé un nouveau financement pour la création d'enregistrements, environ 20000 spécimens de chauves-souris, et nous nous attendons à ce que les fondations que nous avons jeté accélèrent ce travail.

Slide 19

J’aimerais remercier toutes les personnes qui ont contribué par leur temps et leur expertise, pour la désambiguïsation et présents dans le premier paragraphe, mais qui n’ont pas été assignés d’ORCID ID à cet endroit par manque de place. Et merci à la NSF pour le soutien.